МИНИСТЕРСТВА ВЫСШЕГО И СРЕДНЕГО СПЕЦИАЛЬНОГО ОБРАЗОВАНИЯ РЕСПУБЛИКИ УЗБЕКИСТАН

АНДИЖАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

КАФЕДРА РУССКОГО ЯЗЫКА И ЛИТЕРАТУРЫ

ПО ДИСЦИПЛИНЕ

С.С.Камалходжаева

КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

(тексты лекций)

АНДИЖАН

ВВЕДЕНИЕ

В жизни современного общества важную роль играют автоматизированные информационные технологии. С течением времени их значение непрерывно возрастает. Но развитие информационных технологий происходит весьма неравномерно: если современный уровень вычислительной техники и средств связи поражает воображение, то в области смысловой обработки информации успехи значительно скромнее. Эти успехи зависят, прежде всего, от достижений в изучении процессов человеческого мышления, процессов речевого общения между людьми и от умения моделировать эти процессы на ЭВМ.

Когда речь идет о создании перспективных информационных технологий, то проблемы автоматической обработки текстовой информации, представленной на естественных языках, выступают на передний план. Это определяется тем, что мышление человека тесно связано с его языком. Более того, естественный язык является инструментом мышления. Он является также универсальным средством общения между людьми – средством восприятия, накопления, хранения, обработки и передачи информации. Проблемами использования естественного языка в системах автоматической обработки информации занимается наука компьютерная лингвистика. Эта наука возникла сравнительно недавно – на рубеже пятидесятых и шестидесятых годов прошлого столетия. За прошедшие полвека в области компьютерной лингвистики были получены значительные научные и практические результаты: были созданы системы машинного перевода текстов с одних естественных языков на другие, системы автоматизированного поиска информации в текстах, системы автоматического анализа и синтеза устной речи и многие другие.

Данная работа посвящена построению оптимального компьютерного интерфейса средствами компьютерной лингвистики при проведении лингвистических исследований.

Лекция № 1.
Информатизация современного общества.

План:

1. Роль информатики в социальной сфере, в науке, технике, деловом общении.
2. Совершенствование массовой и индивидуальной коммуникации.

Современный человек живет в глобальной информационной среде, а объемы информации возрастают на порядки ежегодно. Современные информационные технологии включают растущее число сетевых коммуникационных технологий, автоматизированных информационных систем, средств массовой коммуникации, систем информационного поиска, систем машинного перевода, а техническая коммуникация развивается в сторону оптимизации информационных процессов и использования удобного для человека языка.

Компьютерная лингвистика, направление в прикладной лингвистике, ориентированное на использование компьютерных инструментов – программ, компьютерных технологий организации и обработки данных – для моделирования функционирования языка в тех или иных условиях, ситуациях, проблемных сферах и т.д., а также вся сфера применения компьютерных моделей языка в лингвистике и смежных дисциплинах. Собственно, только в последнем случае и идет речь о прикладной лингвистике в строгом смысле, поскольку компьютерное моделирование языка может рассматриваться и как сфера приложения информатики и теории программирования к решению задач науки о языке. На практике, однако, к компьютерной лингвистике относят практически все, что связано с использованием компьютеров в языкознании. Проблемы языковой коммуникации «человек – компьютер – человек» и моделирования языка лежат в области исследований такой молодой науки как компьютерная лингвистика (КЛ, Computational Linguistics), которая образовалась на стыке информатики и лингвистики с началом внедрения первых вычислительных машин, а первые эксперименты в этой сфере были связаны с машинным переводом в начале 50-х гг. XX века. Данное направление стало активно разрабатываться в 60–70 гг., и под ним в первую очередь понималось использование статических методов в языкознании, отсюда и название «Computational Linguistics» (в букв.перев. «вычислительная лингвистика»). В России родственные термины «вычислительная лингвистика», «математическая лингвистика» получили распространение в 70-х годах XX века. В конце XX века в связи с развитием компьютерных технологий и их активным приложениям в лингвистических задачах, этот термин как название науки трансформировался, и наука получила более четкое наименование «компьютерная лингвистика». С точки зрения современного подхода основным направлением компьютерной лингвистики является (Natural Language Processing, NLP) задачи АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ЯЗЫКА, что включает задачи анализа и моделирования языковой структуры, а именно:

- графематический/фонематический анализ языка;
- морфологический анализ;
- лексико-грамматический анализ языка;
- синтаксический анализ;
- анализ и моделирование семантической структуры;
- задача синтеза языковых элементов, в т. ч. генерация текстов.

Лингвисты также включают следующие ПРИКЛАДНЫЕ НАПРАВЛЕНИЯ в область интересов компьютерной лингвистики. Мы видим, что большинство нижеследующих задач связано с проблемами автоматической обработки текста:

- машинный перевод;
- распознавание и синтез речи;

- лингвистические основы информационного поиска;
- автоматическое индексирование, реферирование и классификация текстов;
- автоматический контентанализ текста;
- авторизация текстов;
- сетевые технологии представления текста и информации на ЕЯ;
- корпусная лингвистика.

Также к задачам компьютерной лингвистики мы относим:

- разработку и использование искусственных языков, в том числе языков программирования, языков информационных систем;
- компьютерную лексикографию и терминографию;
- компьютерную лингводидактику.

Отдельно в рамках компьютерной лингвистики обычно выделяют задачи и технологии использования статистической лингвистической информации, или автоматическую лингвостатистику.

Таким образом, Компьютерная лингвистика (Computational Linguistics), будучи одним из направлений прикладной лингвистики, изучает лингвистические основы информатики и все аспекты связи языка и мышления, моделирования языка и мышления в компьютерной среде с помощью компьютерных программ, а ее интересы лежат в области:

- оптимизации коммуникации на основе лингвистических знаний;
- создания естественно-языкового интерфейса и технологий понимания языка для общения человека с машиной (это одна из основных проблем Искусственного Интеллекта);
- создания и моделирования информационных компьютерных систем.

Список литературы

1. Лингвистический энциклопедический словарь/ гл. ред. В. Н. Ярцева. – М., 1998. – С. 201-207, 287-289, 615.
2. Свободная энциклопедия языков программирования -http://progopedia.ru/
3. Chomsky, Noam. Syntactic structures. Walter de Gruyter, 2002. – 117 с.
   - books.google.com
4. Герд, А. С. Предмет и основные направления прикладной лингвистки. – http://project.phil.spbu.ru/lib/data/ru/heard/prikling.html
5. Искусственный интеллект// Справочник«Модели и методы».– М., 1990. – Т.2. – С. 7–60.
6. Поспелов, Д. А. Моделирование рассуждений/ Д. А. Поспелов. – М., 1989. – С. 124-151.
7. Кандрашина, Е. Ю. Представление знаний о времени и пространстве в нтеллектуальных системах/ Е. Ю. Кандрашин, Д. А. Поспелов. – М., 1989. – С. 8-34, 248-258.

Лекция № 2.
Ввод языковой информации

План:

1. Искусственные языки как знаковые системы.
2. Языки программирования как искусственные языки.
3. Формальные методы описания искусственных языков. Формальная грамматика. Понятие метаязыка.
4. БНФ-нотации. Синтаксические диаграммы.

1. В предыдущей лекции мы говорили о лингвистических основах информатики и о том, что в связи с невозможностью полной формализации ЕЯ для коммуникации в

компьютерных средах необходима разработка специальных искусственных языков(ИЯ), что входит в сферу интересов лингвистики.

Искусственные языки – это знаковые системы, создаваемые для использования в тех областях науки и техники, где применение естественного языка ограничено, менее эффективно или невозможно.

Любой искусственный язык по сравнению с естественным всегда ограничен (по словарю, синтаксису, семантике слов) и служит для решения определенных задач.

Классификация искусственных языков:

1. Неспециализированные языки общего назначения (например, эсперанто, волапюк);
2. Специализированные языки различного назначения (например, символические языки наук (математика, логика, химия, физика)). Во 2-й класс входят языки человеко-машинного (компьютерного) общения и реализации компьютерных и информационных технологий (языки программирования, языки операционных систем, языки информационных систем и т. п.).

Лингвист должен иметь представление о том, какие бывают языки, какова структурная организация языков и как создаются искусственные языки.

Этому и посвящены следующие наши рассуждения.

2. Языки программирования (ЯП) – это класс искусственных языков, предназначенных для обработки информации с помощью компьютера. Любой язык программирования – это строгая (формальная) знаковая система, при помощи которой записываются компьютерные программы. По разным оценкам, в настоящее время существует от тысячи до десяти тысяч различных языков программирования.

Исторически языки программирования возникли в 40-х годах XX в. и качественно совершенствовались в сторону упрощения своего описания, т.е. высокоуровневой организации, методов программирования и приложения для обработки информации человеком.

Можно выделить следующие качественные уровни развития ЯП (т.е. то, как их классифицируют в программировании):

- Низкий уровень (работа с машинными кодами, например, есть языки–ассемблеры – это версии машинных кодов, адаптированных под аппаратные платформы компьютеров).
- Средний уровень.
- Высокий уровень (язык программирования высокого ровня – это язык, команды и структура которого удобны для восприятия человеком, например, Паскаль, Delphi, SQL, Java Script, PHP и др.)

Если взглянуть на язык программирования как на некий объект с лингвистической точки зрения, можно увидеть, что ЯП так же, как и любой язык, имеет свои ярусы или структуру.

Первый низкий уровень – символьный, его элементы алфавита–буквы, спецсимволы (по аналогии с ЕЯ – графематический уровень).

Второй уровень – это уровень имен, например, зарезервированных слов, выражений (в ЕЯ – это лексический уровень).

Третий уровень – операторный (командный), где синтаксические конструкции имеют повелительный характер (в ЕЯ – аналог синтаксического уровня), и последний – уровень программы, всегда являющейся синтаксически и семантически законченной последовательностью предписаний - команд.

Программа – это структурно строгий текст, записанный по формально заданным правилам искусственного языка программирования.

3. Общим признаком описания специализированных искусственных языков является формальный метод их описания и определения путем задания алфавита, словаря и системы правил образования и преобразования выражений (грамматика). Формальный

метод служит для порождения «правильных выражений» («правильных» – значит «записанных по определенным правилам»).

Например, для языков программирования задаются определенные формы языковых элементов (алфавит, слова), правила построения команд, текстов программ, т.е. точно описываются семантика и синтаксис для однозначного понимания программ компьютером.

Вообще, при написании «правильных выражений», т.е. для формального описания синтаксиса элементов любого языка, широко используются такие нотации как формальные грамматики.

Формальная грамматика – это система строгих (часто математических) правил, позволяющая с помощью единообразных процедур получать (выводить) правильные выражения данного языка либо анализировать имеющиеся выражения на предмет их соответствия правилам языка.

Вопросами формальных грамматик и теорией формальных языков занимается такой раздел языкознания как математическая лингвистика. Она является смежным направлением прикладной лингвистики, тесно соприкасающимся с математикой и информатикой.

В1957 году в своей книге «Syntactic structures» американский ученый-лингвист Ноам Хомский предложил классификацию формальных языков по типу правил формальной грамматики.

Существует множество видов формальных грамматик, как например:

1. Регулярная грамматика.
2. Контекстно-свободная грамматика.
3. Грамматика непосредственно-составляющих.
4. Лексико-функциональная грамматика.
5. Грамматика Монтегю.

Кратко опишем порождающую грамматику.

Порождающая формальная грамматика – это система
G = Vт, Vнт, S, R, где G – грамматика;
Vт – множество терминальных (конечных) символов языка;
Vнт – множество нетерминальных символов (из которых можно выводить далее), заключаются нами в примере ниже в угловые скобки <…>;
S – начальный символ нетерминального множества;
R – система правил вывода типа X,Y
(где X,Y – цепочки символов из Vт, Vнт).
Множество цепочек, выводимых через G из ее начального символа S, есть выражения языка, порождаемые этой грамматикой G (т.е. вывод цепочек всегда начинается с нетерминала S).

Пример:
Формальная система:
Г=<{I, We, They, like, music .},
{S, Pr, V, N}, S, R>,
где{I, We, They, .} – Vт,
{S, Pr, V, N} - Vнт
Система правил R:
<S>□<Pr><V><N>.
<Pr> □I □We□They
<V> □like
<N> □music
Выражения, порождаемые согласно синтаксическим правилам R:

I like music.
We like music.
They like music.
Формальная грамматика, изложенная по подобным правилам, в свою очередь, работает на базе метаязыка, т. е. специальной вспомогательной
системы знаков (нотации) для работы с конечным языком.
IV. На практике применяется еще один метаязык, который даже считают синонимичной записью или функциональным аналогом нотации формальных грамматик. Это Бэкус-Науровы формы (БНФ–формы или БНФ–нотации), которые, как и формальная грамматика, служат для задания правил получения правильных выражений и текстов.
Пример: БНФ-нотация для описания англо-русского словаря:
Пример типовой странички словаря
P
Pay[pei] 1. платить;
2. заработная плата;
3. расплата.
Pea [pi:] горох.
Peak [pi:k] остроконечная вершина.
<словарь> ::= [<раздел>]
<раздел> ::= <заглавная лат.буква>[<словарная статья>]
<заглавная лат.буква> ::= A□B□….□Z
<словарная статья> ::= <термин> <транскрипция>
<перевод>.
<термин> ::= <заглавная лат.буква> [<прописная лат.буква>]
<прописная лат.буква>] ::= a□b□…□z
<транскрипция>::= [[<фонетический знак>]]
< фонетический знак> ::= a:□□□…□z
<перевод> ::= <определение1>□<определение2>
<определение1> ::= <слово>□<словосочетание>
<слово> ::=[<прописная русская буква>]
<прописная русская буква>::= а□б□в….□я
<словосочетание> ::= [<слово> _]
<определение2> ::= 1.<определение1>;
2.<определение1>;
3.<определение1>

Аналогичный метаязык, имеющий графическое наглядное представление – это синтаксические диаграммы, которые используются часто при преподавании программирования на языках высокого уровня. Синтаксическая диаграмма – это схема, объясняющая правило построения либо некоторого элемента, выражения, либо текста.
Пример: Синтаксическая диаграмма морфологической структуры русского слова*:
* Диаграмма представлена нами в ограниченном виде.

КОРЕНЬ СУФФИКС ОКОНЧАНИЕ К Н — О Е
ПРИСТАВКА

Обе эти формы нотаций нашли широкое применение при описании
языков программирования в информатике. Для прикладной лингвистики построение формальных грамматик, БНФ и синтаксических диаграмм интересно как способ понять структуру любого языка, увидеть возможности по моделированию искусственных языков и лингвистических структур.

Список литературы
1. Лингвистический энциклопедический словарь/ гл. ред. В. Н. Ярцева. – М., 1998. – С. 201-207, 287-289, 615.
2. Свободная энциклопедия языков программирования -http://progopedia.ru/
3. Chomsky, Noam. Syntactic structures. Walter de Gruyter, 2002. – 117 с. books.google.com
4. Герд, А. С. Предмет и основные направления прикладной лингвистки. – http://project.phil.spbu.ru/lib/data/ru/heard/prikling.html
5. Искусственный интеллект// Справочник«Модели и методы».– М., 1990. – Т.2. – С. 7–60.
6. Поспелов, Д. А. Моделирование рассуждений/ Д. А. Поспелов. – М., 1989. – С. 124-151.
7. Кандрашина, Е. Ю. Представление знаний о времени и пространстве в нтеллектуальных системах/ Е. Ю. Кандрашин, Д. А. Поспелов. – М., 1989. – С. 8-34, 248-258.

Лекция № 3.
Понятие о лингвистической модели.

План:
1. Моделирование и модель в лингвистике.
2. Модели знаний в искусственном интеллекте. Языки представления знаний как варианты искусственных языков.

1. В разных науках приходится иметь дело с различными моделями (образцами) тех или иных процессов или объектов исследования этих наук.

Метод моделирования – центральный исследовательский метод практически в любой науке, нацеленный на выяснение либо способов, правильности функционирования объекта, либо его свойств при помощи построения модели этого объекта. Метод моделирования языка и языковых процессов широко используется лингвистами, т.к. дает возможность реализовать теоретические знания на практике. Прикладная лингвистика стремится строить модели, отображающие конкретные лингвистические объекты, их системы, а также процессы речемыслительной деятельности человека.

Понятие лингвистической модели также активно используется в компьютерной лингвистике, т. к. без построения лингвистических моделей невозможно решить многие практические задачи создания и использования лингвистических ресурсов (например, машинных переводчиков, электронных
словарей) и задачи автоматической обработки языка.

Всякая модель строится на основе гипотезы о возможном устройстве оригинала и зачастую представляет собой функциональный аналог оригинала, что позволяет переносить полученные знания с модели на оригинал. Критерием адекватности модели часто является эксперимент.

Даже в лингвистике, не говоря о других науках, существует множество определений понятия «модель». Чаще всего под моделью понимают формальный способ представления, описания, лингвистических объектов,
например:

Модель – тип (language pattern) каких-либо устойчивых типовых единиц исследуемого языка (сочетаемостей слов, синтаксических структур).

Такие устойчивые образцы мы неоднократно встречаем в грамматиках при изучении иностранных языков. Например, это стандартная для английского языка

структура предложения, которую называют SVO или Subject-Verb Object pattern (пр. The Big Brother is watching you).

- модель-теория – описание на базе строгой формализованной научной теории, т.е. структура элементов и правил с фиксированным метаязыком (например, теория формальных грамматик).

Таким образом, модель в лингвистике – это формализованная структура или описание с фиксированным метаязыком, служащая образцом для исследования порождаемого языка, анализа его характеристик и функций.

Модель всегда предполагает наличие однозначно заданных объектов, связывающих их отношений и правил обращения с ними (см. пример формальной грамматики в лекции №3).

Понятие лингвистической модели возникло в структурной лингвистике, но вошло в активный научный обиход в 60–70 гг. XX в. с развитием математической лингвистики и проникновением в лингвистику точных формальных методов исследования. Лингвисты обратили особое внимание на семантику единиц языка и на модели речевой деятельности, тесно связанные с моделями мышления (которые часто исследуются в логике). В 70-х годах XX в. в результате неудачных попыток всестороннего полного моделирования естественных языков в рамках решения задач машинного перевода «ученые пришли к выводу, что решение многих прикладных проблем не может быть чисто лингвистическим, а лежит на совсем иных путях, на путях моделирования поведения и мышления человека, семантики, синтеза формальных и семантических средств языка. Так появилась одна из важнейших межотраслевых фундаментальных проблем прикладной направленности– проблема моделирования знаний».

2. Моделирование на базе компьютерных технологий человеческих знаний, понимания языка и человеческого мышления – это задачи Искусственного интеллекта (Artificial Intelligence) как одного из ведущих научных направлений компьютерной науки – информатики, которое занимается созданием интеллектуальных когнитивных технологий и машин, способных понимать, моделировать и анализировать тексты, хранить и перерабатывать естественно-языковую информацию, принимать интеллектуальные решения.

Система «искусственного интеллекта» должна понимать вопросы человека, решать интеллектуальные задачи и вести диалог с человеком на основе заложенных в нее процедурных и декларативных знаний. Примером интеллектуальной искусственной системы является экспертная система, качество которой определяется в первую очередь тем, насколько естественно общение с ней человека при решении задач.

В искусственном интеллекте важно понятие «знания», а при построении интеллектуальных систем – языки представления знаний.

знания, как правило, включают в себя 3 составляющие – опыт, навыки и умения – и бывают декларативными («знаю ЧТО это») и процедурными («знаю КАК это делать»).

Знания в интеллектуальной системе представляют специальными моделями на искусственных языках лингвистического обеспечения этой системы.

Для прикладной лингвистики интересны языки представления знаний и семантики (ЯПЗ). ЯПЗ (языки представления знаний) – это искусственные языки, построенные по законам искусственных языков. Наиболее популярные ЯПЗ(языки представления знаний) – это логические, сетевые, фреймовые и продукционные.

а) Логические ЯПЗ (языки представления знаний) представляют знания в виде синтаксически правильных формул какой-либо формальной логической системы.

Логических систем очень много. Разработаны даже псевдофизические логики, т.е. системы правил, описывающих отношения реального мира, например, логики времени, причины и следствия, логики пространственных отношений. Часто используют логику предикатов и ее язык логических формул.

Примеры логических формул:

1. Вася любит Машу.
Это предложение на ЕЯ можно записать по-другому
Любить (Вася, Маша) или на языке логики предикатов.
L (b, m)
Предикатная формула может быть одноместная и многоместная.
2. Вася не студент. Вася подарил Маше кольцо.
~ S (b) P (b, m, k)
Часто предикатная формула может быть вложенной.
3. Теорема Пифагора на языке логической формулы будет иметь вид:
РАВНЫ[СУММА(КВАДРАТ(1-Й КАТЕТ), КВАДРАТ (2-Й КАТЕТ)),
КВАДРАТ(ГИПОТЕНУЗА)] или на символьном метаязыке P1(P2(P3(k1), P3(k2)), P3(g))
4. В логических формулах часто используются специальные знаки – универсальные кванторы, например, (существует…), (для всех…).
Есть девушки красивее Маши. Клиент всегда прав.
x (G(x) & N(x, m)) x (C(x) →P(x))
Попробуйте подобрать свои примеры про любовь (обозначим ее как L) на естественном языке для таких формальных записей:
x L(x,x) x (L(x,m) →L(x,d)) ~ x y L(y,x) ~ x y (D(y) & L(x,y)
А теперь попробуйте преобразовать следующие два предложения в формальную запись:
Не все студенты ответят на все вопросы.
Вы сможете, если попытаетесь.

Решение задач в логике на логическом ЯПЗ (языки представления знаний) – это логический (дедуктивный) вывод по правилам этой логической системы.

б) Сетевые ЯПЗ (языки представления знаний) в качестве моделей используют семантические сети, где узлы сети – это какие-либо информационные единицы – понятия, факты, процессы, имена, а другие – отношения между ними. Отношения могут быть любыми (временные, причинно-следственные, больше-меньше и т. п.).

Сетевые ЯПЗ (языки представления знаний) часто используют для явного описания отношений в той или иной ситуации или для описания семантики структур.

Решение задач на сетевых моделях сводится к поиску фрагмента сети, совпадающему с данным образцом и к организации логического вывода на семантической сети. Связи элементов семантической сети можно представить в виде формул–записей на метаязыке семантических сетей, т.к. графически можно представить наглядно лишь простые структуры.

в) Фреймовые ЯПЗ (языки представления знаний).

Фрейм (frame – каркас, скелет, рамка) – это шаблон типовой ситуации или некоторая структура, содержащая сведения об определенном объекте, его характеристиках или их значениях и выступающая как целостная единица знаний. Понятие «фрейма» в классическом понимании связано с минимальным описанием факта или явления, у которого нельзя удалить никакую часть описания без того, чтобы не утратить полноту этого описания.

Фрейм может быть ролевой или структурный.

Структурный фрейм в большей мере отражает декларативные знания, т.е. описывает структуру какого-либо понятия, объекта или документа.

Ролевой фрейм, в отличие от структурного, представляет процедурные знания, например, скрипты или сценарии каких либо типовых процедур или работ.

В приложений и стандартных автоматов, например, при оплате услуг в банкоматах или для дистанционных операций. Одна из более сложных областей использования ролевых фреймов– это робототехника.

Основные характеристики этих ЯПЗ (языки представления знаний) – компактность и вложенностью уровней.

г) Продукционные ЯПЗ (языки представления знаний).

Продукции – это одна из популярных форм представления процедурных знаний в экспертных системах и других системах знаний, работающих со сценариями.

Продукции представляют в основном процедурные знания, причем формула правил продукции относительно проста– «ЕСЛИ..., ТО...»:
if<condition> then<conclusion> или if<condition> then <action>
Например,
ЕСЛИ < температура тела более40□>,
ТО <срочно вызывай скорую помощь по телефону03>.
Продукции могут быть составными, например, ifa and b and c thend.

В форме правил-продукций можно задавать и лингвистические знания о языке, полученные в результате анализа множества текстов.
Задачи для продукционной модели ставятся как задачи поиска нужной последовательности продукций, при котором достигается нужная цель.

Список литературы

1. Лингвистический энциклопедический словарь/ гл. ред. В. Н. Ярцева. –М., 1998. – С. 201-207, 287-289, 615.
2. Свободная энциклопедия языков программирования - ttp://progopedia.ru/
3. Chomsky, Noam. Syntactic structures. Walter de Gruyter, 2002. – 117 с. books.google.com
4. Герд, А. С. Предмет и основные направления прикладной лингвистки. – http://project.phil.spbu.ru/lib/data/ru/heard/prikling.html
5. Искусственный интеллект// Справочник«Модели и методы».– М., 1990. – Т.2. – С. 7–60.
6. Поспелов, Д. А. Моделирование рассуждений/ Д. А. Поспелов. – М., 1989. – С. 124-151.
7. Кандрашина, Е. Ю. Представление знаний о времени и пространстве в нтеллектуальных системах/ Е. Ю. Кандрашин, Д. А. Поспелов. – М., 1989. – С. 8-34, 248-258.

Лекция № 4.
Информационный поиск.

План:

1. Информационный поиск.
2. Лингвистика в задачах информационного поиска.
3. Информационно-поисковые языки как искусственные языки.

1. Теория и практика информационного поиска стала развиваться в середине XX века. К концу века с внедрением мировых информационных компьютерных сетей и их информационно-лингвистического обеспечения эта область стала отдельным признанным научно-практическим направлением компьютерной науки со своей теорией информационного поиска и многочисленными практическими разработками, продуктами и технологиями.

Поиск информации (Information Retrieval) – это процесс отыскания в некоторой системе хранения информации таких документов(текстов, записей и т.д.), которые соответствуют поступившему запросу. В качестве таких средств хранения и поиска информации выступают информационно-поисковые системы (IRS/Information Retrieval Systems), элементами которых являются структурированный массив документов (база данных/индекс), выступающих как объект поиска, разнообразные технические и программные средства, как например, программы-роботы, а также информационно-поисковый язык, задающий правила индексирования документов, правила поиска.

По В.П.Захарову «Информационно-поисковая система (ИПС) – это упорядоченная совокупность документов (массивов документов) и информационных технологий, предназначенных для хранения и поиска информации–текстов (документов) или данных (фактов). Информационно-поисковыми системами являются любые определенным образом организованные хранилища информации. Причем информационно-поисковые системы могут быть и неавтоматизированными. Главное – это целевая функция: хранение и поиск информации».

При вводе документа в базу данных ИПС (информационно-поисковые система) его индексируют (поэтому саму базу поисковой системы часто называют индексом). Процесс индексирования в основном связан с определением и выборкой ключевых слов обрабатываемых документов и выражением их формально в виде поискового образа. Так база данных ИПС (информационно-поисковые система) состоит из множества индексных поисковых образов.

Непосредственно при поиске производится сопоставление вашего запроса, т.е. того, что вы указали в запросе с поисковым образом, т.е. тем, что хранится в индексе.

Для более наглядного понимания способа хранения информации в базе ИПС (информационно-поисковая система) приведем следующий простой пример.

В зависимости от объекта хранения и типа запроса часто в учебной литературе различают два вида информационного поиска:

- документальный (когда пользователь ищет какой-либо текст/документ);
- фактографический (когда пользователь ищет какие-либо фактические данные, например, «День рождения Анджелины Джоли»).

Характеристики информационного поиска – это такие его семантические показатели как:

- полнота выдачи информации;
- точность ее выдачи;
- потери информации;
- информационный шум.

Полнотой поиска (Recall) называется мера, вычисляемая как отношение количества выданных релевантных документов к общему числу релевантных документов, содержащихся в базе информационно-поисковой системы.

Точность поиска (Precision) – это отношение количества выданных системой релевантных документов к общему числу документов в выдаче.

Рассмотрим следующую таблицу, по которой легко просчитываются основные показатели информационного поиска:

Документы релевантные нерелевантные (необходимые, нужные) (ненужные)
выданные ИПСa b
невыданные ИПСc d
ПВ= (a/(a+c))·100% – полнота выдачи информации
ПИ= (c/(a+c))·100% – потеря информации
ИШ= (b/(a+b))·100% – информационный шум
ТВ= (a/(a+b))·100% – точность выдачи

Особенно интересно нам понятие «релевантность» как фундаментальное понятие теории информационного поиска. Согласно определению, документ, центральный предмет или тема которого в целом соответствует смысловому содержанию информационного запроса, называется релевантным, а свойство смысловой близости между документом и информационным запросом– релевантностью. По многим, чаще всего субъективным, причинам релевантность была и остается основной проблемой информационного поиска.

2. Теория и практика информационного поиска тесно связана с лингвистикой, т.к., во-первых, основной объем информации и тексты документов представлены на

естественных языках, а для извлечения информации и индексирования текстов необходимы знания автоматической обработки языков, например, правила компьютерной морфологии. Во-вторых, информационные компьютерные системы построены и работают на базе искусственных языков.

Из определения информационно-поисковой системы видно, что ее основным лингвистическим средством является специализированный искусственный язык.

Информационно-поисковый язык (ИПЯ) – это специализированный искусственный язык, предназначенный для:

1) описания информационных запросов к информационно-поисковым системам (язык запросов),

2) описания формальных характеристик документов в виде поискового образа, хранящегося в базе системы (язык индексирования).

Необходимость внедрения искусственных языков вызвана необходимостью устранения избыточности естественного языка для информационного поиска, а также ликвидации языковой синонимии, омонимии и неоднозначностей разного рода.

Информационно-поисковый язык, как и любой язык, состоит из фиксированных единиц, например, имеет свой словарь и синтаксис, и является искусственным языком, т.е. ограниченным по своей форме и структуре стоящими перед ним задачами на поиск.

Рассмотрим каждый из вариантов ИПЯ более подробно:

2.1. Языки запросов

Кажущиеся простыми на первый взгляд языки запросов представляют собой довольно комплексные системы правил и процедур. Как правило, модель языка запросов включает следующие элементы:

- Поисковые единицы (ключевые слова, выражающие информационную потребность пользователя).
- Средства лемматизации лингвистических единиц запроса (приведение слов к нормальной словарной форме).
- Специальные поисковые (булевские) операторы, типа OR, NOT, AND.
- Средства линейной грамматики (операторы расстояния, позиционные операторы, например, title:).
- Дополнительные условия поиска: ограничение области поиска по языку, региону, дате создания документа и т. п.

Запрос на поиск чаще всего ограничен по языку и может иметь определенное формальное представление. Самый частотный способ поискового запроса – через ограниченный набор ключевых слов, задаваемый пользователем в поисковой строке. Например, найти эту книгу в Интернете можно, задав ключевые слова Соснина введение прикладная лингвистика в поисковых системах Яндекс, Google и т. п.

Кроме стандартной практики в поисковых системах предусмотрена функция расширенного языка запросов, ее можно легко найти на сайтах поисковых машин– http://yandex.ru/search/advanced илиhttp://www.google.ru/advanced_search. Расширенный поисковый язык включает множество операторов для улучшения качества и сужения зоны поиска релевантной информации.

Например:

- "чемпионат мира2014" OR"олимпийские игры2014";
- мумий тролль мультфильм -рок -лагутенко – Поиск в Яндексе или Google только мультфильма, а НЕ (-отрицание) группы Ильи Лагутенко;
- социолингвистика-site:ru.wikipedia.org (ищем информацию везде кроме сайта русской Википедии);

- date:ГГГГ{*|ММ{*|ДД}} – Поиск в Яндексе только по страницам, дата которых удовлетворяет заданному условию (например, ищем труды профессора УлГТУ только за 2012 г. Шарафутдинова Н.С. date: 2012).

2.2. Языки индексирования

В литературе по теории информационного поиска [2, 4, 5] обычно выделяют следующие распространенные виды информационно-поисковых языков для индексирования документов:

1. Языки классификаций–иерархические; алфавитно-предметные и др.;
2. Дескрипторные языки.

Иерархические ИПЯ (иерархия – классификация от общего к частному).
Такая организация используется для поиска книг в библиотеке (например, Универсальная десятичная классификация – УДК).
Дерево классов
… … … …

Алфавитно-предметные ИПЯ (информационно-поисковые система) представляются как алфавитный список ключевых слов какого-либо документа с пометами (например, алфавитно-предметный указатель в конце книг). Используется для построения различных указателей, каталогов, картотек.

Дескрипторные языки – наиболее естественная и популярная форма языков индексирования для выражения поисковых образов и в настоящее время широко используется в современных информационно-поисковых системах (в частности во многих поисковых системах сети Internet).

В таких языках используется принцип координатного индексирования, т.е. перечисляются ключевые слова или дескрипторы, которые выражают центральную тему или целостную характеристику искомого объекта. При этом используется принцип логического умножения понятий, в результате которого из простых лексических единиц строятся более сложные, выражающие более узкие понятия. Этот принцип описания содержания документов через перечисление ключевых слов существует издавна. Например, пересечением понятий ПАРИЖ и ДОСТОПРИМЕЧАТЕЛЬНОСТИ, заданных в запросе, порождается новое более узкое понятие ДОСТОПРИМЕЧАТЕЛЬНОСТИ ПАРИЖА.

В качестве лексических единиц в дескрипторных ИПЯ выступают дескрипторы – имена понятий или классов понятий, которые явно перечисляются в дескрипторном словаре. Это слова (или словосочетания), выбранные в качестве представителей классов и групп синонимичных слов и словосочетаний. Как правило, это существительные или номинативные выражения.

Списки дескрипторов организованы в специальные семантические словари (поисковые тезаурусы). Поисковые тезаурусы (одноязычные или многоязычные) – специальные словари для информационного поиска по массивам естественно-языковых документов, организованные по принципу сопоставления слов с их понятиями. Их структура и разработка часто стандартизируются, см., например, ГОСТ7.25-2001.

Такой дескрипторный словарь используется как средство лексического контроля (например, снятия омонимии, синонимии единиц) при индексировании документов и запросов. Например, при индексировании все синонимы запроса и поискового образа представляются одной и той же лексической единицей – дескриптором (ср. лингвистика – языкознание языковедение, наука о языке).

Структурно поисковый тезаурус представляет собой алфавитный список дескрипторов вместе с их словарными статьями (гнездами). Словарная статья обычно содержит:

- заглавный дескриптор;
- ключевые слова или словосочетания, входящие в гнездо данного дескриптора (условные синонимы);

- «вышестоящие» дескрипторы (находящиеся с данным в отношении «род–вид», «часть–целое»);
- «нижестоящие» дескрипторы (находящиеся с заглавным дескриптором в отношении «вид–род», «целое–часть»);
- ассоциативные дескрипторы (связанные с данным другими разнообразными отношениями: причина–следствие, процесс–объект, свойство–носитель свойства, функциональное сходство).

Указанные отношения приводятся с пометами, чаще всего так:

ДЕСКРИПТОР

с– ключевые слова (синонимы)

в– родовые слова (дескриптор, подчиняющий данный)

н– видовые слова (дескриптор, подчиненный данному)

а– ассоциации (отношения)

Нужно отметить, что тезаурусы бывают не только информационно-поисковые, а их структура не ограничивается только представленными отношениями. Тезаурусы относятся к очень интересному классу семантических словарей, что мы рассмотрим в следующей лекции, посвященной одному из древних практических направлений лингвистики–лексикографии.

Список литературы

1. Web information retrieval –http://research.microsoft.com/enus/collaboration/global/asia-pacific/talent/webirhistoryfuturetrends.pdf
2. Захаров В.П. Информационные системы (документальный поиск): учебное пособие/ В. П. Захаров. – СПб. : Изд-во СПбГУ, 2002. – 188 с. – http://vp-zakharov.narod.ru/publications.htm
3. Баранов А.Н. Введение в прикладную лингвистику/А.Н.Баранов. – М. : Эдиториал УРСС, 2001. – 360 с. (стр.197-200)
4. Соколов А.В. Информационно-поисковые системы/ А. В. Соколов. – М., 1981. – стр. 8-13, 40-60, 70-77.
5. Черный А.И. Введение в теорию ИП/ А. И. Черный. – М., 1975. стр. 25-100.
6. ГОСТ 7.25-2001 СИБИД. Тезаурус информационно-поисковый одноязычный. Состав, структура и основные требования к построению. http://www.complexdoc.ru/

Лекция № 5.
Компьютерная лексикография.

План:

1. Лексикография как одно из важных направлений прикладной лингвистики. Традиционная и машинная (компьютерная) лексикография. Основные направления компьютерной лексикографии.
2. Словарь как объект лексикографии. Классификация и организация словарей.
3. Идеографические словари и тезаурусы как примеры словарей.

1. Лексикография (от греч. lexis 'слово' и grafia 'писание, наука'), будучи одним из важных направлений прикладной лингвистики, занимается теорией и практикой составления словарей.

В лексикографии выделяют два научно-практических направления: традиционная и машинная (компьютерная) лексикография.

Традиционная лексикография имеет глубокие исторические корни и в большей мере занимается теорией и практикой составления«традиционных» словарей. Самые ранние глоссы (от греческого glossa 'язык, слово', словарные

пометы о значениях незнакомых слов) известны с глубочайшей древности (например, шумерские глоссы XXV в. до н. э.). Поэтому, самыми первыми типами словарей многие ученые считают глоссарии, написанные от руки
списки иностранных и необычных слов, с которыми приходилось сталкиваться в манускриптах на древних языках.

Машинная лексикография (Computational Lexicography) занимается автоматизацией подготовки словарей и решает задачи разработки электронных словарей. В отличие от традиционной, машинная (компьютерная) лексикография – относительно молодая наука, реализующая традиционные наработки в технических средах и создающая разнообразные электронные словари. К основным задачам компьютерной лексикографии относятся также задачи разработки технологий составления электронных словарей и управления терминологией (Terminology Management).

В настоящее время можно выделить три основных направления компьютерной лексикографии:

1) автоматическое получение из текста различных словарей (например, терминологических, частотных словарей, словарей конкордансов и др.). Примеры страниц частотных словарей см. в Приложении 7;

2) создание словарей, являющихся электронными версиями традиционных словарей (например, словарь Даля), или комплексных электронных лингвистических словарей для традиционных словарных работ, например, известный словарь LINGVO [http://lingvopro.abbyyonline.com/ru]. Большой выбор такого рода открытых словарей доступен через поисковые системы, например, http://slovari.yandex.ru/.

3) разработка теоретических и практических аспектов составления специальных компьютерных словарей, например, для информационного поиска, машинного перевода (например, поисковые тезаурусы, словари стоп-слов, словари основ и флексий для морфологического анализа/синтеза). Словарь стоп - слов для информационного поиска приведен в Приложении 8.

Таким образом, мы видим, что основным объектом интересов лексикографии является такой продукт как словарь и все задачи, связанные с разработкой разного типа словарей.

2. Словарь – определенным образом организованное собрание слов, обычно с приписанными им комментариями, в которых описываются особенности их структуры и/или функционирования. Помимо слов, объектами словарного описания могут выступать их компоненты (таковы, например, словари морфем), словосочетания различных типов, устойчивые группы – пословицы, поговорки, цитаты и т. п.

В другом значении термин «словарь», или лексикон, обозначает всю совокупность слов некоторого языка (иначе говоря, его лексику) и противопоставляется термину «грамматика», обозначающему совокупность правил построения из слов более сложных языковых выражений.

Таким образом, под словарем понимают:

- полный словарный состав языка;
- упорядоченное для решения практических задач множество лексических единиц;
- справочную книгу слов, расположенных в определенном порядке, дающую кому-либо информацию о том или ином слове.

Основная задача словаря – это представление либо описание лексики языка и ее особенностей для решения конкретных задач. Это сложнейшая проблема, так как лексика языка имеет тенденцию увеличиваться и качественно изменяться. Вот как метафорично говорит В.Селегей в статье «Электронные словари и компьютерная лексикография»: «Многие словари, основной корпус статей которых сформировался в языковой атмосфере середины века, представляют собой лексикографические музеи (а то и терминологические кладбища, если говорить о специализированных словарях)».

Основной структурной единицей словаря (как книги) является словарная статья, организация и моделирование которой является зачастую сложнейшей прикладной проблемой и задачей лексикографии.

Словник словаря – это перечень терминов словаря без их толкований.

Важным вопросом при составлении словарей также является порядок расположения словарных статей, чаще всего это алфавитный порядок или предметный (тематический), при котором слова группируются по темам или графически (например, тематический визуальный словарь).

Электронные словари часто представляют из себя сложный комплекс компьютерных программ– лингвистических платформ. Примером тому служит ABBYY Lingvo Content – профессиональное приложение, которое позволяет легко создавать новые словари и глоссарии, редактировать их и публиковать в бумажном и электронном виде на сайтах, на корпоративных порталах, а также в виде приложений для PC, Mac, смартфонов и мобильных устройств (http://www.abbyy.ru/lingvo_content).

Отметим некоторые особенности автоматических словарей:

- кроме словарной базы данных (перечень слов по алфавиту) для автоматической словарной работы необходимы специальные алгоритмы и программы, например, морфологической нормализации, или лемматизации. Лемматизация – это приведение разных форм слова к его канонической (исходной) форме (одна из задач компьютерной морфологии);
- в машинных словарях присутствуют не только перечни отдельных слов, но и до50 % словосочетаний (особенно в терминологических словарях, которые очень важны для перевода специальных и технических текстов);
- электронные словари получили в настоящее время большое распространение в силу их доступности для широкого круга пользователей через различные мобильные и персональные устройства.

3. Мы уже упоминали словарь - тезаурус как тип семантического словаря в предыдущей лекции. Однако тезаурусы имеют более широкие приложения, чем мы указали в лекции, посвященной информационному поиску.

Класс идеографических словарей (предметные, тематические), к которым часто относят и тезаурусы, – это особого рода словари, организованные, во-первых, по тематическому принципу, и, во-вторых, по принципу иерархии отношений и «от смысла к слову», т.е. идеографические словари ориентированы на семантику языка, и каждый такой словарь – это некоторая семантическая модель лексики, построенная на иерархических отношениях типа «род– вид», «часть– целое», «синонимы» и т. п.

Вот что говорится в известном издании про идеографические словари «Альтернативой алфавитному расположению слов является размещение их по смысловой близости. Словари, в которых лексика располагается на основании этого критерия, получили название идеографических (от греч. idea – понятие, идея, образ igrapho – пишу)».

Тезаурус – это также идеографический словарь, но имеющий четкую сложную иерархию отношений. В словарных статьях тезауруса отражены существенные для данного термина связи с другими понятиями, иными словами – это маленькая энциклопедия.

Самый известный классический словарь-тезаурус – это тезаурус П.Роже «Thesaurus of English words and phrases», опубликованный в1852 г., неоднократно переиздававшийся как классический лексикографический труд и представленный в сети Интернет. П.Роже систематизировал лексику английского языка по категориальным

группам, каждая из которых представлена именем понятия, которых сначала насчитывалось 1000. Слова
расположены в алфавитном порядке, далее идут синонимы слов по частям речи (существительные, глаголы, прилагательные, наречия), антонимы и затем списки родственных слов. Многими отмечается, что ценность этого тезауруса в его естественности, в том, что это описание общей понятийной лексики языка, а также в том, что его можно привлекать к использованию в компьютерных системах как семантическое средство.

Направление, связанное с разработкой семантических словарей, активно развивается. Существуют глобальные проекты семантических словарных баз типа WORDNET [9], а также такие любопытные версии словарей как визуальные тезаурусы, наглядно представляющие визуальные карты выбранных слов, например[10]:

Список литературы

1. Энциклопедия Кругосвет – электронная версия – http://krugosvet.ru/enc/gumanitarnye_nauki/lingvistika/SLOVAR.html (дата обращения: 30.08.2012)
2. В.Селегей. Электронные словари и компьютерная лексикография - http://www.lingvoda.ru/transforum/articles/selegey_a1.asp (дата обращения: 30.08.2012)
3. Морковкин В.В. Идеографические словари/В.В.Морковкин. – М.: Изд-во МГУ, 1970. – http://rifmovnik.ru/ideog_book.htm (дата обращения: 30.08.2012)
4. Roget P.M. Thesaurus of English words and phrases. Электронная версия словаря– http://thesaurus.com/Roget-Alpha-Index.html (дата обращения: 30.08.2012)
5. Modern Language Association (MLA): "love." Roget's 21 st Century Thesaurus, Third Edition. Philip Lief Group 2009.14 Aug. 2012. <Thesaurus.com http://thesaurus.com/browse/love> (дата обращения: 30.08.2012)
6. Шарафутдинова Н.С. Лингвокогнитивные основы научно-технической терминологии/ Н.С.Шарафутдинова. – Ульяновск: УлГТУ, 2006.–131 с.
7. Проект экспериментальных словарей – http://dict.ruslang.ru (дата обращения: 30.08.2012)
8. Словарь русского языка: В 4-х т./РАН, Ин-т лингвистич. исследований/ Под ред. А.П.Евгеньевой. – 4-е изд., стер. – М.: Рус.яз.; Полиграфресурсы, 1999. – http://feb-web.ru/feb/mas/mas-abc/default.asp (дата обращения: 30.08.2012)
9. WordNet® – http://wordnet.princeton.edu/ (дата обращения: 30.08.2012)
10. Visual thesaurus – http://www.visualthesaurus.com (дата обращения: 30.08.2012)

Лекция № 6.
Современный машинный перевод

План

1. Машинный перевод. Виды МП.
2. Технологии МП. Задача машинного перевода как одна из важнейших задач прикладной лингвистики. Основные этапы системы МП, работающей на базе правил.

1. Перевод с одного языка на другой в общем случае состоит в изменении алфавита, лексики и синтаксиса языка с сохранением его семантики.

Перевод – это вид информационной деятельности, потребность в которой никогда не сокращается, а наоборот увеличивается. Исследования рынка переводов показали, что объемы этого вида деятельности постоянно увеличиваются, а в составе переводов

преобладают специальные, научно-технические переводы– до половины общего объема рынка переводов, затем
идут устный, учебный, синхронный, …, художественный.

Проблема моделирования перевода для приложения ее в компьютерной среде является центральной проблемой как прикладной лингвистики, так и искусственного интеллекта. Очевидно, что автоматизация перевода позволит повысить его эффективность, а также расширит границы межчеловеческой коммуникации.

Машинный перевод – это преобразование компьютером текста на одном естественном языке в эквивалентный по содержанию текст на другом естественном языке. Вот какое определение мы находим на сайте известной (и старейшей) системы машинного переводаSYSTRAN: «Machine translation(MT) is automated translation. It is the process by which computer software is used to translate a text from one natural language (such as English) to another (such as Spanish). To process any translation, human or automated, the meaning of a text in the original (source) language must be fully restored in the target language, i.e. the translation. While on the surface this seems straightforward, it is far more complex.Translation is not a mere word-for-word substitution. A translator must interpret and analyze all of the elements in the text and know how each word may influence another. This requires extensive expertise in grammar, syntax (sentence structure), semantics (meanings), etc., in the source and target languages, as well as familiarity with each local region» [1].

Системы МП – это целый комплекс специальных сложнейших программ и алгоритмов плюс специальные автоматические словари и формализованные грамматики для каждой языковой пары (входного и выходного языков).

В1954 г. был проведен так называемый Джоржтаунский эксперимент: переводился текст (250 слов) с русского на английский язык. Первая промышленная система МП SYSTRAN начинает функционировать в США в1970 г. Первые опыты разработки коммерческих систем по МП показали огромные трудности при моделировании семантики языковых единиц и разрешении лингвистических неоднозначностей разного рода. Проблемы, связанные с моделированием всей структуры языка, полностью не решены до сих пор, поэтому были предложены другие подходы к организации машинного перевода, работающие на базе информации из больших объемов текстов, например, так появились технологии CAT – Computer-Aided-Translation, а также статистический перевод на базе поисковых систем, например, сервисGoogle Translate, работающий по собственным алгоритмам с 2007 г.

Виды современного МП.

Существует несколько подходов к классификации систем машинного перевода, исходя из разных критериев.
Например, их часто делят в зависимости от степени участия человека в процессе перевода. Согласно этому критерию все системы делятся на три типа:

- FAMT (Fully-Automated Machine Translation) – полностью автоматический машинный перевод;
- HAMT(Human-Aided Machine Translation) – машинный перевод с участием человека;
- MAHT (Machine- Aided Human Translation) – перевод, осуществляемый человеком с привлечением вспомогательных программных и лингвистических средств.

Вот как различают эти термины авторы известного обзора «Survey of the State of the Art in Human Language Technology» концаXX века: «…we would like to take the term machine assisted translation (MAT) as covering all techniques for automating the translation activity. The term human-aided machine translation (HAMT) should be reserved for the techniques which rely on a real automation of the translating function, with some human intervention in pre-edition, post-edition or interaction. The term machine-aided human translation(MAHT) concerns machine aids for translators or revisors…» [6].

Мы также указывали на похожую классификацию в своем предыдущем издании этого пособия:

1) информативный МП– грубый пословный перевод без участия человека, достаточный для поверхностного ознакомления с содержанием текста (автоматический). Активно эксплуатируется и сейчас при поиске информации в поисковых системах;

2) профессиональный МП – более качественный перевод на базе коммерческих систем МП с последующим редактированием человеком (автоматизированный);

3) интерактивный (персональный) МП – считается переводом в специальных системах поддержки, например, в таких как CAT, эксплуатирующих базы перевода (TRANSLATION MEMORY), проходит в режиме диалога человека с компьютерной системой. Активно используется
коммерческими компаниями.

Благодаря постоянному развитию технологий, накоплению лингвистических ресурсов и лингвистической статистики большинство современных систем МП дает неплохое качество перевода. Качество МП во
многом зависит от возможностей настройки, имеющихся ресурсов (например, наличия готовых переводческих баз), от типа текстов(очевидно, что художественные тексты не имеет смысла переводить с помощью систем МП, а тексты инструкций к устройствам переводятся с хорошим качеством даже в автоматическом режиме).

Выше мы указали на одну из классификаций видов машинного перевода.

Сейчас мы рассмотрим то, что часто называют технологиями или подходами к осуществлению машинного перевода. Здесь просматриваются существенные отличия и связано это, на наш взгляд, с тем, что разработчики систем машинного перевода, когда уровень и возможности современных аппаратных средств выросли, пошли по довольно эффективному пути минимизации творческих усилий, что проявилось в создании различных систем памяти переводов и лингвостатистических технологий.

2. Технологии машинного перевода.

Как правило, выделяют следующие технологии, на базе которых работают современные системы машинного перевода:

- Example-based Machine Translation, или Машинный перевод на базе готовых примеров переводов [7].

Эта концепция МП была предложена японским исследователем М.Нагао в 80-х годах XX века. Идея заключалась в следующем при накоплении достаточно большой коллекции ранее переведенных фраз для конкретных типов текстов и в узких областях велика вероятность того, что большая часть последующих текстов будет аналогична уже переведенным вручную.

Ярким представителем этого подхода являются инструменты CAT (Computer Aided-Translation tools), реализующие технологии автоматизированной поддержки перевода.

К системам CAT относят те системы, которые обеспечивают работу на базе «памяти переводов» (Translation Memory), поэтому их часто называют TM-инструменты [8]. Функция обращения к базам «памяти перевода» дает переводчику и компании несомненные преимущества:

- Одинаковую лингвистическую единицу либо предложение, встречающиеся в разных местах, не нужно переводить дважды. Если подобное соответствие переведено, то необходимо лишь откорректировать предыдущий перевод.
- Если перевод текста выполняет группа переводчиков, то законченный текст получается более терминологически и стилистически однородным.

База данных TM состоит из пар параллельно сопоставленных сегментов (чаще всего это предложение), базы формируются программно по ходу
Перевода (изначально CAT - система пустая) и хранятся отдельно. CAT-программа при переводе производит обращение к базе, и если новый сегмент

имеет точное или частичное соответствие в имеющейся накопленной базе переводов, то предлагается вариант перевода, если соответствия нет, то перевод идет вручную и вариант записывается в базу как новый сегмент.

Многими пользователями отмечается, что при использовании CAT -систем значительно повышается эффективность работы переводчика, работающего с типовыми специальными текстами, для которых уже накоплена богатая практика переводческих примеров, т. е. базы переводов.

Самыми популярными в России являются пакеты SDL Trados, Deja Vu.

Кроме них существует большое количество как платных, так и бесплатных версий программ автоматизированного перевода, функционально программы похожи. Выходят демонстрационные бесплатные версии программ. Рынок CAT - программ постоянно развивается, параллельно накапливается множество специальных баз переводов.

- Statistical Machine Translation, или Статистический МП.

Эта технология машинного перевода также эксплуатирует идею вышеописанного EBMT - подхода, но имеет более строгую математическую базу, учитывая статистику тех лингвистических закономерностей, которые получены на базе структурного анализа текстов и анализа доступных параллельных корпусов текстов [7]. Корпус параллельных текстов– это тексты, содержащие предложения на одном языке и соответствующие им предложения на втором(например, двуязычные сайты). Статистический МП использует свойство «самообучения» языку (machine learning), т. е. чем больше накоплено параллельных текстов и чем точнее они соответствуют друг другу, тем лучше результат статистического машинного перевода. Считается, что интерес к использованию лингвостатистических методов в МП возник еще в конце40-х годовXX века параллельно с развитием теории информации К.Шеннона. В 60-х годах XX века исследователи успешно применяли лингвостатистику к решению задач дешифровки древних текстов.

Но активно это направление стало развиваться только в начале 90-х годов XX века во время технологического и информационного бума, накоплением данных в корпусной лингвистике, машинном обучении и информационном поиске. Поэтому глобальные поисковые системы стали внедрять сервис статистического перевода в свои интерфейсы. Например, поисковик Google перешел на эту технологию в2007 году и предлагает сервис Google Translate.

В начале 2011 года Яндекс, как ранее Google, внедрил собственную подобную систему машинного перевода.

О принципах статистической технологии очень удачно и доходчиво изложено на сайте Яндекса. Приведем лишь выбранные отрывки в учебных целях (т.к. часто тексты полезных статей имеют тенденцию пропадать со временем на просторах Интернета):

«Такой перевод основывается не на правилах языка (системе эти правила даже не известны), а на статистике. Чтобы выучить язык, система сравнивает сотни тысяч параллельных текстов... Это могут быть, например, большие тексты с разноязычных версий сайтов организаций…

В системе машинного перевода Яндекса три основные части: модель перевода, модель языка и декодер.

Модель перевода— это таблица, в которой для всех известных системе слов и фраз на одном языке перечислены все возможные их переводы на другой язык и указана вероятность этих переводов (для каждой пары языков есть своя таблица). Модель перевода создается в три этапа: сначала подбираются параллельные документы, потом в них– пары предложений, а затем уже пары слов или словосочетаний. … Система сравнивает не только отдельные слова, но и словосочетания из двух, трёх, четырёх или

пяти слов, идущих подряд. В переводчике Яндекса модель перевода для каждой пары языков содержит более миллиарда пар слов и словосочетаний.
Другая составляющая системы машинного перевода — модель языка. Для её создания система изучает сотни тысяч различных текстов на нужном языке и составляет список всех употребленных в них слов и словосочетаний с указанием частоты их использования. Это знание системы о языке, на который нужно перевести текст.
Непосредственно переводом занимается декодер. Для каждого предложения исходного текста он подбирает все варианты перевода, сочетая между собой фразы из модели перевода, и сортирует их по убыванию вероятности. Например, пользователь захотел перевести фразу «to be or not to be». Допустим, из всех вариантов в модели перевода максимальная вероятность получилась у сочетания «быть или не бывает», сочетание «быть или не быть» оказалось с небольшим отрывом на втором месте и так далее. Все получившиеся варианты сочетаний декодер оценивает с помощью модели языка. В данном примере модель языка подскажет декодеру, что «быть или не быть» употребляется чаще, чем «быть или не бывает».
В итоге декодер выбирает предложение с наилучшим сочетанием вероятности (с точки зрения модели перевода) и частоты употребления (с точки зрения модели языка)».
Подробнее: http://company.yandex.ru/technologies/translation/

- Rule-based Machine Translation, или Машинный перевод на базе лингвистических правил, – это классическая технология МП, резко отличающаяся от предыдущих, с разработки которой и началась история МП и компьютерной лингвистики.

Глобальной задачей такой технологии машинного перевода является формальное моделирование естественных языков как объектов, реализованных в определенной программной и аппаратной компьютерной среде, а также моделирование различных процедур анализа, сопоставления (трансфера) и синтеза лингвистических единиц и текстов для этих естественных языков. т.е. такая задача требует реализации процедур:
1) графематического (символьного) анализа;
2) морфологического и лексического анализа и синтеза;
3) синтаксического анализа и синтеза;
4) семантических преобразований.
При МП обрабатываются все уровни языковой структуры, переходя от одной подзадачи к другой. Для моделирования МП необходимо моделирование его подзадач, разработка алгоритмов и компьютерных программ для работы этих алгоритмов. Если внимательно присмотреться к этим задачам МП, то можно увидеть, что они представляют отдельные и сложные проблемы и направления прикладной лингвистики.
За историю машинного перевода как научного направления выделились три подхода к моделированию систем такого рода [8, 9]:
а) «прямые» системы МП давали пословный перевод (слово в слово);
б) системы «интерлингвы», или «текст–смысл–текст», целью было разработать смысловой язык-посредник;

Первый подход ограничен по своим возможностям, а для моделей-интерлингвы пока недостаточно научной базы и успешных разработок.
в) «текст–трансфер-текст». Перевод происходит на уровне переводных соответствий языковой пары. Единицей перевода выступает переводное соответствие, тесно связанное с лексической единицей (словосочетание, слово). В этой модели есть промежуточный этап– трансфер, на котором происходит установление переводных соответствий (согласование языковых единиц).
Преобразование Т1 должно начинаться с предварительной подготовки–его анализа. Обычно различают следующие виды лингвистического анализа: морфологический, лексический, синтаксический, семантический. Целью этапа анализа является получение

всей лингвистической информации о выделенных единицах данного текста, например, получение всей морфологической информации о словоформе для последующего корректного перевода, и построение внутреннего представления входного предложения.
На этапе трансфера уже начинается переводческая работа, происходит сопоставление единиц по словарям, выполняются иные преобразования с помощью специальных формальных правил трансфера.
Цель этапа синтеза на основе полученной в результате анализа информации построить (синтезировать) правильное предложение на выходном языке.
Эта модель построения систем перевода на базе правил наиболее удачная в коммерческом плане и в настоящее время преобладает над прямыми и интерлингвальными типами МП(на ней работают известные промышленные системы PROMT и SYSTRAN).
Мы кратко описали лишь базовые понятия и технологии современного машинного перевода, но это направление постоянно развивается, появляются новые способы автоматизации и повышения эффективности переводческой деятельности, различные гибридные технологии. Например, появляются технологии типа Traduwiki (http://traduwiki.org) для совместной переводческой деятельности, т.н. «облачные технологии» (cloud-based translation), а люди продолжают придерживаться любимого ими, уже отработанного веками принципа«principle of least effort».

Список литературы

1. SYSTRAN – сайт системы http://www.systran.co.uk/systran/corporateprofile/translation-technology/what-is-machine-translation
2. Краткая история машинного перевода. Журнал «Русский репортер» –
3. http://rusrep.ru/2010/24/istoriya_perevoda/ http://en.wikipedia.org/wiki/History_of_machine_translation Yorick Wilks . Machine Translation: Its Scope and Limits. – Springer, 2008 – 254 p. - http://books.google.com/
4. William John Hutchins. Machine translation: past, present, future. – Ellis Horwood, 1986. – 382p.
5. Survey of the State of the Art in Human Language Technology –http://www.lt-world.org/hlt-survey/master.pdf (п. 8.3.1.)
6. Harold Somers. Review Article: Example - based Machine Translation/ http://kitt.cl.uzh.ch/clab/satzaehnlichkeit/tutorial/Unterlagen/Somers1999.pdf
7. Семенов, А. Л. Современные информационные технологии и перевод:
8. учеб. пособие/ А.Л.Семенов. – М.: Академия, 2008. – 224 с.
9. Марчук Ю.Н. Методы моделирования перевода/ Ю.Н.Марчук. – М., 1985. – С. 7-17, 43-45, 91-110, 135-137.

Лекция № 10.

Перспективы развития компьютерной лингвистики.

План:

1. Текст и гипертекст. Задачи автоматической обработки текста.
2. Современная лингводидактика.
3. Ближайшие и отдаленные задачи. Связь развития компьютерной лингвистики с прогрессом в компьютерных науках.

Кратко рассмотрим другие прикладные задачи лингвистики, которые появились либо получили новое развитие с возникновением компьютерной техники.

С появлением компьютерных сетей и новых информационных технологий несколько трансформировались традиционные лингвистические понятия, в частности, понятие текста.

Под текстом (лат. Textus – связанность, материал, сплетение), письменным или устным, принято понимать логически связанную последовательность лингвистических знаков. Основные характеристики текста: связанность;
осмысленность; цельность (текст должен быть закончен по смыслу). Кроме
того, классически текст имеет линейную структуру.

С середины 80-х годов XX века, когда быстрыми темпами стали развиваться компьютерные и телекоммуникационные сети (WWW, Internet) и
вместе с тем глобальная коммуникация, в научный обиход вошло новое понятие гипертекста (Hypertext) [1]. Оказалось, что для решения многих информационных задач нелинейный и мультимедийный характер текста может быть эффективнее. Гипертекст – это технология организации информации и особым образом структурированный текст, разбитый на отдельные блоки, имеющий нелинейное представление, для эффективной презентации информации в компьютерных средах. Для стандарта представления информации в сетях стали разрабатываться специализированные языки.
Например, HTML – Hyper Text Markup Language – язык гипертекстовой разметки документов. При работе с текстом можно решать множество задач от общефилологических до статистических. Условно направление обработки текста с помощью доступных компьютерных и информационных технологий можно назвать автоматической обработкой текста, т.е. это любое преобразование текста на естественном или искусственном языках с помощью компьютера. Автоматическая обработка текстов осуществляется, как мы заметили раньше, в технологиях информационного поиска (при индексировании текстов), машинного перевода, словарной работе. В этом
разделе мы рассмотрим некоторые неупомянутые ранее приложения и задачи.

- Автоматизированные операции по печати, редактированию и верстке текста Первоначально пользователю были доступны примитивные операции по обработке текста в простых компьютерных программах-редакторах(Lexicon), впоследствии требования к редактированию и представлению документов возрастали, что привело к созданию усовершенствованных систем типа Word Processors (Microsoft Word и др.). Сейчас это развитые программы с функциями обработки таблиц, графики, проверки орфографии и стилистики. Кроме того, есть специальные издательские системы для профессиональной верстки документов, газет, рекламной продукции, книг (например, Page Maker или продвинутые пакеты типа Adobe® Digital Publishing Suite).
- Распознавание текста

Распознавание текста в этом смысле привязано к задаче распознавания символов (Character Recognition) и реализуется эффективнее для печатного текста, чем для рукописного. С работой лингвистического распознавателя пользователь сталкивается при сканировании текста, например, через платформу ABBYY Fine Reader. В

настоящее время задача распознавания печатного текста практически решена в силу ограниченного набора печатных гарнитур. Для рукописного текста(технологию называютICR, т. е. Intelligent Character Recognition) распознавание графем гораздо сложнее, так как нужны более совершенные алгоритмы распознавания образов.

- Автоматическое реферирование и аннотирование текстов

Автоматическое реферирование (Automatic Text Summarization) – это составление коротких изложений материалов, аннотаций или дайджестов, т.е. извлечение наиболее важных сведений из одного или нескольких текстовых документов и генерация краткого содержания анализируемого объекта.

Аннотирование, в отличие от реферирования, предполагает еще большее сжатие содержания исходного текста. Существует много способов решения этой задачи прикладной лингвистики. Как правило, автоматическое реферирование основано на выборке ключевых фрагментов документов – выделении наиболее информативных фраз и элементов– и их сборке.

При автореферировании активно используется статистическая информация по информативности разных элементов текста по частоте их появления в нем, информативность элемента текста также зависит от его позиции в документе или текстовых маркеров(например, термины выделены курсивом или жирно), которые характеризуют их содержательную значимость. Существуют как коммерческие, так и открытые, например, Open Text Summarizer [2], отдельные продукты для реферирования, а также встроенные функции автореферата (например, вMS Word).

- Корпусная лингвистика

Корпусная лингвистика (Corpus Linguistics) – целое направление прикладной лингвистики, активно развивающееся с конца XX века, а в России – с начала XXI века, например, Национальный корпус русского языка, стал разрабатываться лишь в 2004 году[3]. В настоящее время электронных корпусов текстов для разных языков и приложений существует большое множество.

Задачи корпусной лингвистики связаны с разработкой технологий построения электронных лингвистических ресурсов особого типа– корпусов текстов (Corpora) – и решением задач разного рода на базе этих текстов.

В основном, такие коллекции и массивы текстов отражают реальное функционирование того или иного языка, а их перенос в компьютерные среды активизировал их практическое и широкое использование в прикладной лингвистике.

Корпусная лингвистика дает материал для различного рода исследований языка и его вариантов и определяет основной метод анализа текстов и языка на базе корпусов. Одной из важных особенностей метода анализа на базе корпусов является исследование не только чисто лингвистических явлений (грамматических или лексических функций слов, их связей с другими лексемами), но и таких явлений, как, например, частотности лексем или грамматических конструкций в тех или иных жанрах, диалектах.

Корпусный подход, или метод лингвистического исследования, основанный на корпусах текстов, ориентирован на прикладное изучение языка, его функционирования в реальных средах и текстах, что важно, например, для преподавания языка и для компьютерной лингводидактики, что затрагивается нами в последнем разделе курса.

II. Современная лингводидактика

В первой лекции мы уже обращали с вами внимание на то, что англоязычный термин «Applied Linguistics» за рубежом часто трактуется несколько уже, чем мы привыкли в России, и под ним понимается направление, связанное с методами и технологиями преподавания языка, т.е. лингводидактика. Этот взгляд на сферу интересов прикладной лингвистики существует и у нас в стране и поддерживается Национальным обществом прикладной лингвистики, НОПРИЛ, штаб которого находится в Московском государственном университете и руководит которым профессор С.Г.Тер-Минасова [5].

Одним из современных направлений прикладной лингвистики является возможность автоматизированного обучения иностранным языкам и поддержки его преподавания. За рубежом это направление, известное как ComputerAssisted Language Learning and Teaching (CALL/CALT), является перспективным в силу многих объективных причин и преподается как специальная дисциплина прикладного языкознания на педагогических и лингвистических факультетах колледжей и университетов.

Применение компьютерных мультимедийных, дистанционных и открытых технологий в процессе обучения иностранным языкам положительно зарекомендовало себя в последние десятилетия [6]. Мультимедиа технологии позволяют моделировать среду, имитирующую лингвистическую и коммуникативную реальность, что очень важно для языкового обучения, а также активизировать основные методические принципы обучения иностранным языкам– развитие навыков: аудирования, говорения, чтения и письма. Неотъемлемой частью компьютерного образования с внедрением новых информационных и сетевых технологий стали такие электронные средства его поддержки, как машинные переводчики, словари, национальные корпусы текстов. Огромный потенциал для CALL/CALT несет доступ в Internet, в котором можно найти миллионы различных ресурсов для повышения эффективности иноязычного образования, а также различные технологии дистанционной коммуникации, например, Skype.

Список литературы

1. Потапова, Р. К. Новые информационные технологии и лингвистика: учебное пособие/ Р. К. Потапова– М. : МГЛУ, 2002. – 576 с. Open Text Summarizer – http://libots.sourceforge.net/
2. Национальный корпус русского языка - http://www.ruscorpora.ru
3. Британский национальный корпус – www.natcorp.ox.ac.uk
4. Национальное общество прикладной лингвистики –http://www.nopril.ru/
5. Ken Beatty. Teaching and Researching Computer-Assisted Language Learning (разделы- A brief history of CALL, Hypertext hypermedia and multimedia, Eight CALL applications) – books.google.com

ЗАКЛЮЧЕНИЕ

Спецификой нашего учебного пособия является ориентация студентов на профиль практической деятельности будущих молодых специалистов в области теоретической и прикладной лингвистики. Курс лекций изначально дает установку на прикладной характер лингвистических исследований и сразу определяет круг основных задач этой предметной области. Рассмотренные в данном учебном издании задачи и сферы применения практических знаний специалистов не ограничены указанными, напротив, прикладная лингвистика является живой, интересной областью знаний, которая пополняется новыми лингвистическими приложениями и технологиями параллельно с развитием информационного общества.

В заключение укажем, что по окончании вуза выпускники профиля теоретическая и прикладная лингвистика, как показывает опыт, могут применить полученные знания информационных технологий, иностранных языков и компьютерной лингвистики и работать в туристических и информационных агентствах, образовательных структурах, переводческих фирмах, в редакционных и информационных отделах организаций в качестве:

- лингвистов-переводчиков;
- разработчиков технической документации (технических писателей);
- специалистов информационных служб;
- специалистов по интернационализации и локализации изделий на международных рынках;

SEO - оптимизаторов, разработчиков WEB-сайтов коммерческих фирм на иностранных языках.

На получение такого рода профессиональных компетенций и направлен наш профиль обучения.

СПИСОК ОБЯЗАТЕЛЬНОЙ ЛИТЕРАТУРЫ И ЭЛЕКТРОННЫХ РЕСУРСОВ

1. Баранов А.Н. Введение в прикладную лингвистику: учебное пособие/ А.Н.Баранов. – 2-е изд., испр. – М. : Едиториал УРСС, 2009. – 360 с.
2. Большой энциклопедический словарь. Языкознание / под ред. В.Н.Ярцевой – М.: Большая Российская энциклопедия, 1998. – 685 с. (см. соответствующие разделы и определения под терминами– социолингвистика, психолингвистика и т. д.).
3. Герд А.С. Предмет и основные направления прикладной лингвистки. – http://project.phil.spbu.ru/lib/data/ru/heard/prikling.html (дата обращения: 08.08.2012)
4. Захаров В.П. Информационные системы (документальный поиск): учебное пособие/ В.П.Захаров. – СПб.: Изд-во СПбГУ, 2002. – 188 с. – http://vpzakharov.narod.ru/publications.htm (дата обращения: 11.08.2012)
5. Морковкин В.В. Идеографические словари/В.В.Морковкин. – М.: Изд-во МГУ, 1970. – http://rifmovnik.ru/ideog_book.htm (дата обращения:17.08.2012)
6. Селегей В. Электронные словари и компьютерная лексикография- http://www.lingvoda.ru/transforum/articles/selegey_a1.asp (дата обращения: 17.08.2012)
7. Семенов А.Л. Современные информационные технологии и перевод: учеб. пос. / А. Л. Семенов. – М. : Академия, 2008. – 224 с.
8. Потапова Р.К. Новые информационные технологии и лингвистика: учеб. пос. / Р. К. Потапова. – М. : МГЛУ, 2002. – 576 с.
9. Шарафутдинова Н.С. Лингвокогнитивные основы научно-технической терминологии/ Н.С.Шарафутдинова. – Ульяновск: УлГТУ, 2006. – 131 с.
10. Шарафутдинова Н.С. Теория и история лингвистической науки: учебник/ Н.С.Шарафутдинова.– 3-е изд., испр. и доп. –Ульяновск: УлГТУ, 2012. –139с.
11. Этимологический словарь – Online Etymological Ductionary – http://www.etymonline.com/index.php?term=linguist (дата обращения: 23.08.2012)
12. Chomsky, Noam. Syntactic structures. Walter de Gruyter, 2002 – 117 с. – books.google.com (дата обращения: 30.08.2012)
13. Daniel Chandler. Semiotics for beginners (разделIntroduction) – http://www.aber.ac.uk/media/Documents/S4B/sem01.html (дата обращения: 30.08.2012)
14. Robert Marty. 76 Definitions ofThe Sign by C.S.Peirce – http://www.cspeirce.com/rsources/76defs/76defs.htm (дата обращения: 30.08.2012)
15. Roget P.M. Thesaurus of English words and phrases. Электронная версия словаря – http://thesaurus.com/Roget-Alpha-Index.html (дата обращения: 23.08.2012)
16. Survey of the State of the Art in Human Language Technology – http://www.ltworld.org/hlt-survey/master.pdf (дата обращения: 19.07.2012)
17. Harold Somers. Review Article: Example-based Machine Translation/ http://kitt.cl.uzh.ch/clab/satzaehnlichkeit/tutorial/Unterlagen/Somers1999.pdf (дата обращения: 28.07.2012)
18. Ken Beatty. Teaching and Researching Computer-Assisted Language Learning (разделы – A brief history of CALL, Hypertext hypermedia and multimedia, Eight CALL applications)- books.google.com (дата обращения: 28.07.2012)

ДОПОЛНИТЕЛЬНАЯ РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА ПО КУРСУ:

1. Герд А.С. Прикладная лингвистика/ А.С.Герд. – СПб.: Изд-во С.-Петерб. ун-та, 2005. – 268 с.
2. Звегинцев В.А. Мысли о лингвистике/ В.А. Звегинцев / серия «Из лингвистического наследия Звегинцева», 2008. – 336 с.
3. Зубов А.В. Информационные технологии в лингвистике: учеб. пособие/ А. В. Зубов. – М.: Академия, 2004. – 206 с. (Высшее профессиональное

образование. Языкознание).
4. Искусственный интеллект: справочник: в3 кн. / под ред. Э. В. Попова. – М.: Радио и связь, 1990. – Кн. 1: Системы общения и экспертные системы. – 464 с.
5. Искусственный интеллект: справочник: в3 кн. / под ред. Д. А. Поспелова. – М.: Радио и связь, 1990. – Кн. 2: Модели и методы. – 303 с.
6. Кибрик А. Е. Очерки по общим и прикладным вопросам языкознания (универсальное, типовое и специфичное в языке) / А. Е. Кибрик. – 3-е изд., стер. – М. : УРСС, 2002. – (Лингвистическое наследие20 века). – 336 с.
7. Кандрашина Е.Ю. Представление знаний о времени и пространстве в интеллектуальных системах/ под ред. Д. А. Поспелова. – М. : Наука, 1989. – (Проблемы искусственного интеллекта; вып. 16). – 328 с. С. 8-34, 248-258.
8. Леонтьева Н.Н. Автоматическое понимание текстов: учебное пособие/ Н. Н.Леонтьева. – М. : Академия, 2006. – 304 с.
9. Марчук Ю.Н. Методы моделирования перевода/ Ю.Н.Марчук. – М., 1985. С. 7-17, 43-45, 91-110, 135-137.
10. Марчук Ю.Н. Компьютерная лингвистика: учеб. пособие для студ. вузов, специализирующихся по направлению и спец. «Филология» / Ю. Н. Марчук. – М. : АСТ: Восток-Запад, 2007. – (Языкознание). – 317 с.
11. Марчук Ю.Н. Основы терминографии: метод. пособие/ Ю.Н.Марчук. – М.: ЦИИ МГУ, 1992. – 76 с.
12. Поспелов Д.А. Моделирование рассуждений/ Д.А.Поспелов. – М., 1989. – С. 124-151.
13. Потапова Р.К. Речь: коммуникация, информация, кибернетика: учеб. пособие для вузов по спец. «Автоматизированные системы обработки информации и управления», «Лингвистика» / Р. К. Потапова. – М. : Радио и связь, 1997, 2001, 2010 – 528 с.
14. Прикладное языкознание: учебник/Л.В.Бондарко и др.–СПб.,1996.–528 с.
15. Соколов А.В. Информационно-поисковые системы/А.В.Соколов.–М., 1981. – С. 8-13, 40-60, 70-77.
16. Черный А.И. Введение в теорию ИП/А.И.Черный. – М., 1975. – С. 25-100.
Дополнительные электронные ресурсы (часть ресурсов также указана в тексте пособия):
17. www.dialog-21.ru/ (общая информация по направлениям КОЛИНТ)
18. Сайт кафедры ОТИПЛ МГУ(общая информация) http://www.philol.msu.ru/~otipl/new/main/index.php
19. Открытая электронная энциклопедия«Википедия» (запросы по темам курса), например, http://en.wikipedia.org/wiki/Linguistics, http://ru.wikipedia.org/wiki/Linguistics
20. Открытая электронная энциклопедия «Кругосвет» (запросы по темам курса) - http://www.krugosvet.ru
21. Свободная энциклопедия языков программирования- http://progopedia.ru/
22. Пирс Ч. Прагматизм. Научная метафизика. - http://filosof.historic.ru/books/item/f00/s00/z0000699/st007.shtml
23. ГОСТ 7.25-2001 СИБИД. Тезаурус информационно-поисковый одноязычный. Состав, структура и основные требования к построению. http://www.complexdoc.ru/
24. ГОСТ Р7.24-2007 СИБИД. Тезаурус информационно-поисковый многоязычный. Состав, структура и основные требования к построению. – 2008 г. http://www.lib.tsu.ru/index_main.php?id=11
25. Проект экспериментальных словарей– http://dict.ruslang.ru
26. SYSTRAN – сайт системыhttp://www.systran.co.uk
27. Краткая история машинного перевода. Журнал «Русский репортер» – http://rusrep.ru/2010/24/istoriya_perevoda/
28. http://en.wikipedia.org/wiki/History_of_machine_translation

29. Yorick Wilks . Machine Translation: Its Scope and Limits. – Springer, 2008 – 254 p. – http://books.google.com/
30. William John Hutchins. Machine translation: past, present, future. – Ellis Horwood, 1986. – 382p.
31. Claire Bradin Siskin. CALL on the Web – http://edvista.com/claire/call.html
32. Open Text Summarizer – http://libots.sourceforge.net/
33. Национальный корпус русского языка– http://www.ruscorpora.ru
34. Британский национальный корпус– www.natcorp.ox.ac.uk
35. Национальное общество прикладной лингвистики– http://www.nopril.ru/
36. Фундаментальная электронная библиотека «Русская литература и фольклор» (ФЭБ) – http://feb-web.ru/
Дополнительные материалы по курсу также представлены в электронном виде на сервере лаборатории 304 в каталоге1 курса и предоставляются по запросу студента.