Белорусский государственный университет
Факультет социокультурных коммуникаций
Кафедра компьютерной лингвистики и лингводидактики

СОГЛАСОВАНО
Заведующий кафедрой
_____О.Г. Прохоренко
протокол № __ от __.__.2017

СОГЛАСОВАНО
Декан факультета
_____И.И. Янушевич
протокол № __ от __.__.2017

Регистрационный номер ___ от ___.___.2017

УЧЕБНО-МЕТОДИЧЕСКИЙ КОМПЛЕКС ПО УЧЕБНОЙ ДИСЦИПЛИНЕ

Машинный перевод

для специальности 1-21 06 01-01 «Современные иностранные языки (преподавание)»

В.В. Воронович

Рассмотрен и утвержден
на заседании Научно-методического совета БГУ
«__» _____ 2017 г., протокол № ___

Минск
2017

Решение о депонировании вынес
Совет факультета социокультурных коммуникаций
Протокол №___ от «___» _____

Составитель:
***В.В. Воронович,* ст. преподаватель кафедры компьютерной лингвистики и лингводидактики факультета социокультурных коммуникаций Белорусского государственного университета**

Рецензенты:

Б. М. Лобанов, доктор технических наук; гл.научн.сотр. Лаборатории распознавания и синтеза речи ОИПИ НАН Беларуси

А.О. Долгова, кандидат филологических наук, доцент кафедры компьютерной лингвистики и лингводидактики ***Белорусского государственного университета***

Воронович, В.В. Машинный перевод: учебно-методический комплекс для специальности 1-21 06 01-01 «Современные иностранные языки (преподавание)» / В.В. Воронович; БГУ, Фак. социокультурных коммуникаций, каф. компьютерной лингвистики и лингводидактики. - Минск: БГУ, 2017. - 57 с.

Учебно-методический комплекс предназначен для студентов специальности 1-21 06 01-01 «Современные иностранные языки (преподавание)», выпускаемых факультетом социокультурных коммуникаций БГУ. В структуру учебно-методического комплекса входит учебная программа дисциплины, курс лекций, тематика семинарских занятий и КСР, вопросы к экзамену, библиографический список. В курсе лекций изложены основные теоретические положения о принципах и методах разработки систем машинного перевода, об их лингвистическом обеспечении, об истории разработок и стратегиях построения существующих систем.

ОГЛАВЛЕНИЕ

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Основная цель курса «Машинный перевод» состоит в формировании у студентов знаний о принципах и методах разработки систем машинного перевода, об их лингвистическом обеспечении, об истории разработок и существующих системах, а также в формировании умений и навыков формализации знаний о языке и алгоритмизации лингвистического анализа с целью создания лингвистического процессора систем машинного перевода.

Задачи курса:

- ознакомить студентов с теоретическими основами, методами и средствами формализации языка;
- выработать умения составления лингвистических алгоритмов анализа слова и его значения, структуры и семантики предложения и текста при создании лингвистического процессора системы машинного перевода;
- сформировать навыки использования систем машинного перевода.

Курс «Машинный перевод» тесно связан с изучаемыми на отделении СИЯ (преподавание) дисциплинами «Компьютерная лингвистика», «Автоматическая обработка естественного языка», «Проблемы искусственного интеллекта», «Квантитативная лингвистика», «Инженерия знаний»; комплекс названных дисциплин призван сформировать целостное представление о современных исследованиях в области искусственного интеллекта.

Курс предполагает углубленное знакомство с современными направлениями использования языка в информационных технологиях; с формализацией знаний о языке и алгоритмизацией лингвистического анализа; с принципами создания лингвистических банков данных и баз знаний. Приложение знаний специализации может найти применение в преподавании иностранного языка с использованием новых информационных технологий в школе и вузе; в создании систем машинного перевода; в создании лингвистических банков данных и научной работе на базе этих банков данных.

Курс «Машинный перевод» состоит из цикла лекций и цикла практических занятий. Лекции носят проблемный характер, связаны с наиболее важными направлениями исследований и призваны предоставить студентам знания в предметной области. Практические занятия направлены на формирование умений и навыков создания лингвистических банков данных, лингвистического процессора систем машинного перевода, формализации знаний о естественном языке.

I ТЕОРЕТИЧЕСКИЙ РАЗДЕЛ

1.1 Тексты лекций (фрагменты)

ЛЕКЦИЯ 1

Машинный перевод как направление искусственного интеллекта

1. *Определения машинного перевода.*
2. *Краткие сведения из истории развития направления.*
3. *Стимулы к развитию исследований.*
4. *Преимущества машинного перевода.*

1. *Определения машинного перевода.*

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.

2. *Краткие сведения из истории развития направления.*

Датой рождения машинного перевода как научного направления принято считать 1946 г., когда Уоррен Уивер, директор отделения естественных наук Рокфеллеровского фонда, в переписке с Эндрю Бутом и Норбертом Винером впервые сформулировал концепцию машинного перевода, которую несколько позже (в 1949 г.) развил в своем меморандуме “Translation”, адресованном фонду.

В 1952 г. состоялась первая конференция по МП в Массачусетском технологическом университете, а в 1954 г в Нью-Йорке была представлена первая система МП — IBM Mark II, разработанная компанией IBM

совместно с Джоржтаунским университетом (это событие вошло в историю как Джорджтаунский эксперимент).

К началу 50-х годов целый ряд исследовательских групп в США и в Европе работали в области МП. В эти исследования были вложены значительные средства, однако результаты очень скоро разочаровали инвесторов. Одной из главных причин невысокого качества МП в те годы были ограниченные возможности аппаратных средств: малый объем памяти при медленном доступе к содержащейся в ней информации, невозможность полноценного использования языков программирования высокого уровня.

С развитием вычислительной техники в конце 70-х годов (появление микрокомпьютеров, развитие сетей, увеличение ресурсов памяти) машинный перевод вошел в эпоху "Ренессанса". При этом несколько сместились акценты: исследователи теперь ставили целью развитие "реалистических" систем МП, предполагавших участие человека на различных стадиях процесса перевода.

90-е годы принесли с собой бурное развитие рынка персональных компьютеров (от настольных до карманных) и информационных технологий, широкое использование сети Интернет (которая становится все более интернациональной и многоязыкой). Все это сделало возможным, а главное востребованным, дальнейшее развитие систем МП.

3. *Стимулы к развитию исследований.*

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.

4. 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.

ЛЕКЦИЯ 2

Виды и стратегии машинного перевода

1. *Классификация переводов.*
2. *Типы машинного перевода по степени автоматизации.*
3. *Стратегии машинного перевода.*

1. *Классификация переводов.*

Существуют две основные классификации видов перевода:

a) по характеру переводимых текстов (жанрово-стилистические особенности оригинала);

b) по характеру речевых действий переводчика в процессе перевода (формы перевода).

Жанрово-стилистическая классификация переводов в зависимости от жанрово-стилистических особенностей оригинала выделяет три функциональных вида перевода:

- художественный;
- общественно-политический;
- специальный.

Объектом **художественного** перевода являются художественные произведения. Основная задача любого художественного произведения заключается в достижении образно-эмоционального и эстетического воздействия на читателя. В целях достижения определенного эстетического воздействия на переводной язык используется огромное количество разнообразных языковых средств, от эпитета (красочное определение) до ритмико-синтаксического построения фразы. Такая эстетическая направленность отличает художественный перевод от остальных видов перевода. Машинному переводу по указанным причинам художественные тексты поддаются труднее всего.

Общественно-политическим переводом называется перевод текстов общественно-политического и публицистического характера с пропагандисткой или агитационной установкой. Общественно-политический перевод характеризуется яркой эмоциональной окраской с большой насыщенностью различной терминологии.

Специальный перевод обслуживает различные предметные отрасли знаний, имеющие специфическую терминологическую номенклатуру. Объектом специального перевода являются материалы, которые относятся к различным сферам человеческого знания и практики науки и техники. Эти материалы характеризуются предельно точным выражением мысли, следовательно, широким использованием терминологии. Такой перевод проще всего реализуем в машинном виде, поскольку специальные тексты характеризуются строгой логикой построения текста, жесткой структурой предложений, отсутствием эмоциональной окраски и подтекста.

Под **формами** перевода понимается способ, при котором осуществляется перевод:

– письменный (письменно-письменный, зрительно-письменный, письменный перевод на слух);

– устный (устный перевод на слух, зрительно-устный перевод или перевод с листа, т. е. устный перевод зрительно воспринятого исходного письменного текста).

Машинной реализации проще поддается письменный перевод, поскольку устный требует решения дополнительной задачи – распознавания и синтеза устной речи.

2. Типы машинного перевода по степени автоматизации.

В настоящее время выделяют **типы машинного перевода по степени автоматизации**:

– полностью автоматический;

– автоматизированный машинный перевод при участии человека (с пред-, интер- или постредактированием);

– перевод, осуществляемый человеком, с использованием компьютера (например, с использованием электронных словарей).

3. Стратегии машинного перевода.

Первые системы МП характеризуются **стратегией прямого (пословного) перевода**. Сущность этого подхода к построению МП заключается в том, что исходный текст на входном языке постепенно через ряд этапов преобразуется в текст выходного языка. Преобразования сводятся к тому, что слово (словосочетание) на входном языке заменяется на его словарный эквивалент на выходном языке. Понятно, что в системах первого поколения, использующих стратегию прямого перевода, нет необходимости моделировать функционирование языковой системы в

целом. Для работы таких систем оказывается вполне достаточно правил словарных соответствий. В редких случаях проводится анализ контекста для перевода неоднословных выражений, опять-таки представленных в словаре системы. Важно иметь в виду, что стратегия прямого перевода не делает различий между пониманием (анализом) и синтезом (порождением), поскольку они фактически исключены из преобразований по правилам словарных соответствий. Прямой перевод всегда привязан к конкретной паре языков. Например, неоднозначность выражений входного языка разрешается только в той степени, в которой это оказывается необходимым для выходного языка. По временным рамкам системы первого поколения в основном создавались в период с конца 40-х до середины 60-х гг.

Существенная модификация стратегии прямого перевода обнаруживается в системах с **трансфером** — этапом межъязыковых операций, не сводимых только к замене лексем входного языка на словарные соответствия выходного языка. Наличие этапа трансфера предполагает построение «промежуточного» или «внутреннего» представления, которое далее «приспосабливается» к структуре предложения выходного языка. В отличие от первой стратегии, в архитектуре систем МП с трансфером анализ (понимание) и синтез существуют как особые процедуры и обслуживаются различными алгоритмами.

Развитие идеи трансфера привело к появлению **перевода, основанного на глубинном лингвистическом анализе.** Данная стратегия подразумевает анализ входного текста на всех языковых уровнях (морфологическом, синтаксическом, семантическом, прагматическом), а также многоуровневый синтез выходного текста.

Критика стратегии прямого перевода привела к созданию стратегии **языка-посредника (интерлингвы)**. Главная особенность этой стратегии заключаются в том, что между структурами входного языка и структурами выходного языка находится один или несколько промежуточных языков, на которые по соответствующим правилам последовательно «переписываются» выражения входного языка. Анализ и синтез при использовании языка-посредника принципиально разделяются. Анализ ведется в категориях входного языка, а синтез — в категориях выходного. В качестве языка (языков)-посредников могут выступать языки представления синтаксической и семантико-синтаксической структуры, чисто семантические языки, языки глубинной семантики,

приближающиеся к концептуальному представлению в категориях теории знаний (фреймов, сценариев, планов).

В последнее время получила развитие стратегия **памяти переводов**. Память переводов — база данных, содержащая набор ранее переведенных текстов. Одна запись в такой базе данных соответствует «единице перевода», за которую обычно принимается одно предложение (реже — часть сложносочинённого предложения, либо абзац). Если очередное предложение исходного текста в точности совпадает с предложением, хранящимся в базе, оно может быть автоматически подставлено в перевод. Новое предложение может также слегка отличаться от хранящегося в базе. Такое предложение может быть также подставлено в перевод, но переводчик будет должен внести необходимые изменения.

ЛЕКЦИЯ 3

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.*

Шаг 4. *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.*

Шаг 2. Задача разбиения текста на слова и предложения. Несмотря на кажущуюся простоту, задача разбиения текста на слова и предложения в общем случае далеко не тривиальна. Основную сложность представляют различные сокращения, инициалы, прямая речь, слова, написанные через дефис и т.д.

Распознавание слов ведется с помощью специальных шаблонов. Данные шаблоны описывают различные буквенные, цифровые и буквенно-цифровые группы и символы пунктуации, которые затем будут выделяться в качестве отдельных слов. Например, как отдельные слова будут выделены даты, записанные в их цифровом выражении, номера параграфов и подпараграфов, сокращений вместе с точками (по специальному словарю), а также слова, написанные через дефис в случае, если они распознаются специальным модулем словаря – модулем анализа сложных слов. Этот модуль служит для распознавания словосочетаний типа *черно-белый*, *пряно-острый*. Анализ и перевод таких слов осуществляется на основе специальных правил морфологических преобразований прилагательных. В результате анализа выделенных слов, некоторым словам (инициалам, сокращениям и т.п.) будут присвоены специальные маркеры, которые позволят разрешить многозначность при распознавании границ предложений. Также на этом этапе происходит нормализация слов с целью подготовки их для поиска по словарю.

Шаг 3. Морфологический анализ. Решение данной задачи базируется на словаре исходного языка. В результате поиска по словарю каждому слову предложения приписывается множество лексико-грамматических классов: часть речи, падеж, число, род, категория и т.д., что позволяет в дальнейшем производить сравнение классов, основанное на определенных характеристиках (например, проверять согласование прилагательных и существительных). Процесс поиска слов по словарю предполагает, кроме поиска оригинального слова в случае, если оно не было найдено в словаре, поиск слов с удалением возможных префиксов. Для эффективного поиска префиксов используется древовидная структура, элементами которой являются буквы предлогов. Поиск останавливается либо когда нет дальнейшего перехода в дереве, либо когда найден предлог и слово без этого предлога существует в словаре. Кроме словаря предлогов, для каждого из языков существует таблица межъязыкового соответствия, с помощью которой на этапе синтеза текста получается результирующее слово. На этапе распознавания классов производится также выделение

словосочетаний, которые, согласно словарю, переводятся одним словом (словарь идиом): *по барабану, зайти в тупик, kick the bucket (*дословно – *пнуть ведро, 'сыграть в ящик')*. Далее считается, что все такие словосочетания представляются одним словом. Это гарантирует правильность согласования и перевода словосочетания как единого целого.

Шаг 4. Синтаксический анализ. Сначала для каждого слова производится поиск главного слова, с которым оно должно быть согласовано в результате перевода. При этом не предполагается, что уже обязательно должна быть полностью снята многозначность. В процессе поиска главных слов производится основное снятие многозначности.

Построение синтаксического дерева производится путем последовательного распознавания заранее заданных лингвистических шаблонов и применения на их основе определенных операций.

Основными операциями при распознавании шаблонов являются:

а) проверка, является ли слово определенной частью речи с конкретными характеристиками (например, является ли слово существительным в родительном падеже);

б) проверка, является ли некоторое слово омонимом, т.е. может ли оно принадлежать к разным частям речи (например, *жаркое, love*);

в) проверка согласования двух слов (полное согласование – прилагательное-существительное, согласование по падежу – существительное-существительное и т.д.): *красивая девушка, Татьяна Васильева*;

г) получение семантических характеристик управления предлогов и глаголов – каждый глагол и предлог требуют наличия соответствующих падежей у определяемых ими слов (управляют конкретным падежом). Набор этих падежей зависит от смысла этих предлогов и глаголов (например, *находиться* управляет только предложным падежом, а *писать* управляет дательным, винительным, творительным и предложным падежами).

В случае если какой-либо шаблон распознан, возможны следующие операции над элементами, которые он покрывает:

– удаление из списка лексико-грамматических классов слов всех классов, которые не удовлетворяют определенным условиям (например, удаление всех классов, кроме классов, имеющих среди своих характеристик именительный падеж);

– фиксирование зависимости между словами (например, глагол-существительное в именительном падеже);

– удаление слова из множества отдельных слов предложения и добавление его к одному из главных слов с указанием типа зависимости, т.е. слово становится зависимым и удаляется из рассмотрения при последующих шагах построения дерева для упрощения правил анализа.

Шаг 5. Семантический анализ исходного текста. Основная задача данного этапа – разрешение многозначности на основе полученного дерева зависимостей. Для этого первоначально производится разрешение многозначности базовых слов. Как показали исследования, целесообразным является попарное согласование рядом стоящих базовых слов в порядке, обратном положению слов в предложении. После того, как всем базовым словам поставлен в соответствие один лексико-грамматический класс, производится “досогласование” зависимых от них слов. Параметры выбора лексико-грамматических классов зависимых слов выбираются согласно типу зависимости и лексико-грамматическому классу главного слова.

Пример правила для снятия многозначности:

Обработка омонимии «прилагательное-причастие». Если перед омонимом прилагательное-причастие стоит запятая, то у рассматриваемого слова удаляются словоформы прилагательного. В противном случае удаляются словоформы причастия: *послышался странный звук, свистящий словно ветер; с – звук свистящий.*

Шаг 6. Осуществление перевода построенного дерева. Процесс перевода состоит из следующих шагов:

а) производится пословный перевод базовых слов дерева (кроме глаголов) зависимостей с сохранением оригинального лексико-грамматического класса (либо наиболее близкого к нему по своим характеристикам класса);

б) для глаголов из списка базовых слов, имеющих в качестве исходных характеристик признак рода, перевод осуществляется в множество глаголов одной парадигмы с признаками единственного числа и рода; для остальных глагольных форм – перевод осуществляется с сохранением исходных лексико-грамматических характеристик;

в) для зависимых слов результатом перевода является множество слов, которое определяется на основе типа зависимости и лексико-

грамматических классов главных слов (например, результатом перевода прилагательных является вся парадигма, большинство местоимений переводятся с сохранением исходных падежей) – окончательные лексические характеристики определяются на этапе синтеза.

Также на этом шаге производится анализ слов, результатом перевода которых является словосочетание. В результате перевода словосочетание должно быть согласованным. Для этого, при обнаружении данной ситуации, производится достраивание дерева зависимостей на основе главных слов словосочетаний.

Шаг 7. Осуществление согласования переведенного дерева. В результате перевода получается частично согласованное дерево зависимостей. Для получения полного согласования достаточно использовать процедуру, аналогичную процедуре окончательного разрешения многозначности, применяемую на этапе построения дерева. Так как перевод осуществлялся на основе дерева зависимостей, то данная процедура позволит получить согласованное представление предложения на результирующем языке.

Далее на основе дерева производится построение результирующего предложения. Для этого для каждого слова в словаре результирующего языка производится поиск с целью получения конкретной словоформы, соответствующей зафиксированному лексико-грамматическому классу. Также производится дополнение получившихся слов переводами приставок, если они были удалены из исходного слова при анализе.

ЛЕКЦИЯ 4

Структура систем машинного перевода

1. *Вспомогательные программные средства.*
2. *Состав лингвистической базы данных.*
3. *Лингвистический процессор.*

1. *Вспомогательные программные средства.*

На этапе обработки исходного текста в системе машинного перевода должны присутствовать определенные вспомогательные программные средства. Чтобы преобразовывать тексты из одного кодового

представления в другое, необходимы **конверторы**. Типичный конвертор поддерживает подмножество кодовых таблиц, используемых для символов конкретного языка. Так, для русского и белорусского актуальны кодировки CP866, CP1251, Unicode.

При обработке документов возникает задача преобразования исходного документа в обычный текст (plain text), для дальнейшей его семантической и синтаксической обработки. При этом на этапе преобразования необходимо сохранить полезную информацию о структуре документа и его стилевом оформлении, о взаимосвязи между абзацами, о заголовках и т.д. В задачу **преформатора** входит распознавание различных форматов документов поступающих на вход и выделение текстовой информации из этих документов с сохранением ее структуры. Поскольку существует большое количество различных форматов документов, то возникает необходимость написания преформаторов для каждого из этих форматов. Такая задача является очень трудоемкой и, кроме того, с появлением нового формата документов, возникает необходимость написания отдельного преформатора под новый формат. Чтобы избежать подобных проблем, на практике, как правило, используют преформаторы, преобразующие различные форматы документов к одному, как правило, наиболее легко поддающемуся структуризации. Обычно в качестве такого формата выбирается HTML, XML и т.д. Далее создается модуль анализа уже одного, выбранного в качестве базового, формата, на выходе которого получается структурированный текст с рядом выбранных для заданной глубины анализа текста преформатором признаков. При появлении же новых форматов появляется необходимость лишь в создании преформатора, преобразующего этот формат документов в базовый.

Отдельным блоком можно выделить **программы взаимодействия с базами данных,** которые необходимы для полноценного лингвистического анализа исходного текста и синтеза текста на переводном языке.

Для удобной работы с системой машинного перевода необходимо обеспечить **дружественный интерфейс**, оперирующий доступным для любого пользователя языком управления заданиями. Именно с помощью этого интерфейса пользователь должен иметь возможность без программирования ввести, отредактировать и проверить текст, сделать его перевод, вывести его на печать, провести статистическую обработку материала, получить нужную справку и т. п.

2. Состав лингвистической базы данных.

Лингвистическая база данных для системы машинного перевода включает в себя накопленные лингвистические данные, объективированные текстами, картотеками, словарями, грамматиками и другими лингвистическими источниками. Типичный состав лингвистической базы данных можно ограничить следующими компонентами:

а) **Лексико-грамматический классификатор** свойств исходного языка и переводного языка (система морфологического кодирования). При анализе исходного текста каждое слово в нем должно получить соответствующие морфологические характеристики: признак части речи, род, падеж, наклонение, число и др. Система кодирования должна быть единой для конкретной системы машинного перевода.

б) **Базовый двуязычный морфологический словарь**. В этом словаре устанавливается пословное соответствие каждой словоформы исходного языка словоформам языка перевода.

в) **Словарь сокращений и аббревиатур**. Словарь используется на этапе разбиения исходного текста на слова и предложения. Сокращения и аббревиатуры должны быть по возможности расшифрованы, так как они могут являться членами предложения, следовательно, их необходимо учитывать при синтаксическом и семантическом анализе.

г) **Словарь идиом**. Данный словарь применяется до синтаксического анализа, поскольку очень часто идиома является одним членом предложения и рассматривается как единое целое; при переводе идиома на исходном языке может соответствовать одному слову на переводном языке.

д) **Терминологические словари по предметным областям**. Дополнительные словари подключаются при необходимости перевода специализированных текстов.

е) **Синтаксический словарь**. В данном словаре должна содержаться информация о синтаксической сочетаемости членов предложения как в языке оригинала, так и в переводном языке, а также синтаксические соответствия, необходимые при переводе.

ж) **Семантический словарь** (тезаурус, онтология). Данный компонент содержит информацию о семантической сочетаемости лексем, о

лексико-семантических полях, применяется на этапе построения семантического графа предложения.

з) **Корпус параллельных текстов**. Корпус содержит тексты на языке оригинала и их переводы на другой язык. При нахождении предложения или его фрагмента в корпусе параллельных текстов в текст перевода вставляется его соответствие на переводном языке. На использовании корпуса текстов построена технология памяти переводов.

3. Лингвистический процессор.

Лингвистический процессор предназначен для полного лингвистического анализа текста на исходном языке, а также синтеза текста на языке перевода. Лингвистический процессор включает следующие компоненты:

а) Программа разбиения текста на предложения и слова.

б) Программа распознавания устойчивых словосочетаний. Идиомы должны анализироваться и переводиться как неделимое целое.

в) Программа расшифровки сокращений и аббревиатур.

г) Программа морфологического аннотирования исходного текста.

д) Программа синтаксического анализа и построения дерева зависимостей.

е) Программа семантического анализа и построения семантического графа каждого предложения исходного текста.

ж) Программа выбора переводного соответствия из двуязычного словаря или корпуса параллельных текстов.

з) Программа семантического синтеза текста на переводном языке.

и) Программа построения синтаксической структуры предложения и определения порядка слов в синтезируемом предложении.

к) Программа морфологического синтеза словоформ в переведенном тексте.

ЛЕКЦИЯ 5

Лингвистические проблемы машинного перевода

1. *Проблема многозначности при машинном переводе.*
2. *Синтаксические трансформации в машинном переводе.*
3. *Перевод фразеологических сочетаний в системах машинного перевода.*

1. *Проблема многозначности при машинном переводе.*

Проблема разрешения лексической **многозначности** является одной из самых сложных прикладных задач, связанных с лексическим значением. Задача автоматического (реже полуавтоматического) разрешения лексической многозначности была впервые сформулирована в рамках направления науки и технологии, связанного с созданием систем машинного перевода. В дальнейшем проблема разрешения лексической многозначности стала одной из ключевых не только при создании систем машинного перевода, но и систем обработки текстов на естественном языке других назначений (поиск, классификация).

Полисемия (от греч. polysemos – многозначный) (многозначность) – наличие у единицы языка более одного значения – двух или нескольких. Часто, когда говорят о полисемии, имеют в виду многозначность слов как единиц лексики. **Лексическая полисемия** – способность одного слова служить для обозначения разных предметов и явлений действительности. **Грамматическая полисемия** – совпадение разных грамматических форм одной лексемы.

Реализацию того или иного значения слова осуществляет контекст или ситуация, общая тематика речи. Точно так же, как контекст обусловливает конкретное значение многозначного слова, в определенных условиях он может создавать семантическую диффузность, то есть совместимость отдельных лексических значений, когда их разграничение не осуществляется (и не представляется необходимым). Некоторые значения проявляются только в сочетании с определяющим словом; в некоторых сочетаниях значение многозначного слова представлено как фразеологически связанное. Не только лексическая сочетаемость и словообразовательные особенности характеризуют различные значения

слов, но также в ряде случаев и особенности грамматической сочетаемости.

В прикладных задачах компьютерной лингвистики не делается различия между омонимичными и полисемичными значениями слова. Это связано с тем, что в подавляющем большинстве прикладных задач важна не столько этимология слова, сколько его семантика. Распознавание и разделение групп омонимичных значений также входит в задачу разрешения лексической многозначности, так как иногда может оказаться полезным с практической точки зрения.

Проблема многозначности считается решенной, если для слова выбрано его регулярное значение или если найден синонимический эквивалент в виде регулярного значения для метафорического использования.

Известно, что при разрешении многозначности существует ряд самостоятельных задач. В частности, можно выделить наиболее крупные, «классические» задачи:

1) Задача приписывания известного значения известной лексеме.

2) Задача приписывания известного значения новой лексеме.

3) Задача выявления нового значения для известной лексемы.

4) Задача выявления нового значения для новой лексемы.

Отдельно можно рассмотреть задачи, которые пока сравнительно редко рассматриваются в теории многозначности, но имеют актуальность в прикладных задачах:

1) Задача идентификации имени собственного и отнесения его к онтологическому классу.

2) Задача идентификации использования слова в переносном значении (метафора, метонимия, синекдоха).

Если известно, что должно быть на входе (слово) и что на выходе (значение) системы разрешения многозначности, то разработка системы сводится к созданию и наполнению словарей слов и значений, а также к разработке механизмов разрешения многозначности.

Для новых слов, включая имена собственные, необходимо сначала сформулировать перечень возможных значений, а затем перейти к решению классической задачи разрешения многозначности.

Различают два основных класса механизмов разрешения многозначности.

1 класс. Это механизмы автоматические, предполагающие полностью компьютерное решение этой задачи.

2 класс. Это механизмы интерактивные (диалоговые, полуавтоматические), предполагающие совместное решение задачи человеком и компьютером, и сводятся к тому, что компьютер предоставляет пользователю набор альтернатив, из которого он должен выбрать один вариант.

Одним из автоматических методов разрешения многозначности являются **фильтры**, то есть методы, не выявляющие точного значения, но в явном виде накладывающие ограничения на их спектр. Примерами такого рода фильтров являются правила сочетаемости лексем, правила входимости актантов в синтаксемы и предикативные структуры.

К механизмам разрешения многозначности относятся и те, которые не используют лексического значения, постулированного в явном виде, как это сделано, например, в толковом словаре. Такие методы, как правило, носят **статистический** характер.

Для перевода многозначных слов также используются **контекстологические словари**, словарные статьи которых представляют собой алгоритмы запроса к контексту на наличие или отсутствие контекстных определителей значения. Для каждого многозначного слова указывается его приоритетный переводной эквивалент, специфичный для рассматриваемой предметной области. В настоящее время нет необходимости соединять контекстологический словарь и наборы контекстов со специальной алгоритмической процедурой, поскольку современные языки программирования дают возможность разнообразной реализации системы словаря на компьютере в зависимости от общих условий его функционирования.

При интерактивном методе автор (редактор) текста составляет с помощью опорного толкового словаря родного языка смысловые дополнения, а переводы слов, словосочетаний с учетом дополнений осуществляются с помощью специальных словарей исходного и целевых языков, согласованных со словарем.

В опорном толковом словаре исходного языка совмещаются функции толкового словаря и переводного словаря. Данный словарь отражает те

элементы исходного языка, которые имеют особое значение при переводе хотя бы на один из целевых языков, входящих в систему согласованных (переводных) словарей данного исходного языка. Значения представлены в виде отдельной секции, следующей вслед за описаниями тех смысловых значений слова, для которой они являются общими. Это позволяет учитывать в процессе кодирования многообразие не только лексических, но и грамматических значений.

Процесс смыслового кодирования исходного текста выполняется в компьютере автора исходного текста с помощью служебной программы, содержащей упомянутый опорный толковый словарь исходного языка и реализующей по указаниям автора операции формирования смысловых дополнений. В процессе кодирования автор анализирует последовательно, слово за словом, исходный текст и выделяет очередное слово особым шрифтом в случае, если, по мнению автора (в некоторых случаях – по инициативе служебной программы), данное слово обладает хотя бы одним из следующих признаков:

а) данное слово является многозначным, причём его сочетание с соседними словами может не содержать информации, достаточной для выбора смыслового значения, наиболее близкого к исходному тексту;

б) грамматическая форма данного слова и связанных с ним слов не отражает тот или иной оттенок фактического смысла текста, хотя в переводе на целевой язык данное слово и (или) связанные с ним слова могут иметь конкретные грамматические формы, выбор которых строго зависит от контекста;

в) присутствуют глаголы, причастия, деепричастия, форма которых в исходном языке не отражает однозначно тот или иной характер описываемого в тексте действия и (или) состояния, достигнутого в результате действия, в то время как в том или ином целевом языке для выражения указанных оттенков действий и (или) состояний используются, в зависимости от фактического смысла текста, глаголы, причастия, деепричастия, имеющие конкретные грамматические формы;

г) данное слово вместе с некоторыми соседними словами представляет собой словосочетание, для перевода которого может потребоваться поиск среди известных словосочетаний, относящихся к данному слову, причем в некоторых случаях возможны различия в лексическом составе или в структуре, не влияющие на иносказательное значение словосочетания, например, имеются «вклинившиеся» слова, в

частности, определения или обстоятельства к тем или иным словам, уточняющие значение словосочетания в целом, или вводные слова, добавляются переменные компоненты к началу или концу, изменяются те или иные собственные слова или их последовательность; в связи с этим возникает проблема определить, что некоторые слова принадлежат к сочетанию, найти границы внутри фразы, определить ведущее (ключевое) слово и, наконец, выбрать значение, соответствующее контексту.

Далее служебная программа вызывает из опорного толкового словаря словарную статью, соответствующую отмеченному автором слову, затем автор поясняет смысл этого слова, сопоставляя исходный текст с теми или иными элементами статьи.

Служебная программа обладает некоторыми инициативными функциями, например, указывает автору на несовпадение употребления слова или словосочетания в исходном тексте и в отмеченном элементе словарной статьи, а также указывает автору на слова, пропущенные в процессе анализа исходного текста, но, возможно, обладающие той или иной многозначностью.

2. Синтаксические трансформации в машинном переводе.

Глагольно-именные трансформации – центральный вопрос формирования структуры переводного высказывания. Новое содержание проблеме языковых трансформаций придают современные реалии: необходимость проектировать и развивать обучающие компоненты систем машинного перевода и обработки текстовых знаний на основе уже существующих и вновь создающихся корпусов параллельных текстов.

На современном этапе лингвистических исследований и разработок необходимо синергетическое сочетание функционального и уровневого подходов. Функциональный подход интегрирует языковые средства (синтаксические, лексические, словообразовательные и словоизменительные), принадлежащие разным уровням языка, на основе их функционально-семантических характеристик.

Каждая лексическая форма связана с грамматическими формами в двух направлениях. С одной стороны, лексическая форма, даже когда она взята сама по себе, абстрагировано, обнаруживает значимую грамматическую структуру. С другой стороны, лексическая форма в любом конкретном высказывании, являясь особой языковой формой, всегда сопровождается той или иной грамматической формой. Она выступает в определенной функции, и случаи в которых преимущественно

данная лексическая форма встречается, составляют в совокупности ее грамматическую функцию. Лексические формы, выполняющие какие-либо общие функции, принадлежат к одному формальному классу. На основе различных функций могут возникать частично совпадающие формальные классы. Так, выполнение функции действующего лица характерно для субстантивных выражений и для типично инфинитивных словосочетаний.

Под трансформациями понимаются, прежде всего, преобразования предикаторов в имена и имен в предикаторы: *бежать – бег, учитель – учительствовать*, при этом сохраняется частичное тождество формы – корень или основа слова и определенное тождество семантики. Трансформации постоянно выступают как одно из двух главных средств – наряду с перифразами – создания высказываний.

Отсутствие полного совпадения между языковыми конструкциями в разных языках можно обнаружить при изучении сравнительной частоты употребления в них отдельных частей речи, что важно для построения систем машинного перевода.

Функционально-семантический подход, исследующий отношения «функциональной синонимии» разнородных и разноуровневых единиц языка, чрезвычайно актуален в настоящий момент, когда проводятся эксперименты по выявлению изофункциональных и изосемичных языковых структур из параллельных текстовых корпусов. Именно этот подход позволяет найти соответствия в текстах на разных языках. В самом деле, заранее нельзя с полной достоверностью определить, каким именно образом была переведена та или иная языковая структура в текстовом корпусе. Поэтому необходимо строить и исследовать различные гипотезы при проектировании лингвистического процессора. Функции реализуются при взаимодействии языковых объектов и их контекстов.

а) Трансформации «глагол – имя»

Для научного изложения в целом характерен признак номинативности, т.е. более широкое использование существительных, чем в других функциональных стилях. При этом сопоставительный анализ переводов показывает, что, например, в русском языке эта тенденция выражена более четко, чем в английском, и при переводе английские глаголы нередко заменяются существительными.

Наиболее продуктивные типы глагольно-именных трансформаций при англо-русском переводе коррелируют со следующими функциональными значениями.

1. Обстоятельства цели и следствия, выраженные инфинитивом.

2. Составное сказуемое с инфинитивом.

3. Адъективные трансформации инфинитива.

4. Инфинитив в функции второго дополнения.

5. Инфинитив, стоящий в начале предложения и выполняющий функцию подлежащего.

б) Трансформации «имя – глагол»

Перевод английского герундия на русский язык вызывает затруднения, связанные с его двойственной природой - это неличная форма глагола, выполняющая в предложении функции, исконно присущие имени существительному: подлежащего, дополнения; а также функции определения и обстоятельства, свойственные, соответственно, имени прилагательному и наречию. Морфологически герундий совпадает с действительным причастием английского языка, которое также может играть роль определения и обстоятельства, но не может быть ни подлежащим, ни дополнением в предложении. Эта ситуация является неиссякаемым источником ошибок при переводе даже для человека-переводчика (это одна из центральных тем в курсе теории и практики перевода), а в существующих системах машинного перевода различение форм причастия и герундия вообще не происходит, отдельные виды конструкций реализованы лишь фрагментарно.

Система правил трансфера для машинного перевода вначале строится про принципу одновариантных правил, когда переводное соответствие подбиралось как наиболее широкий способ перевода некоторой конструкции, пусть не всегда совершенно грамматичный, однако же обеспечивающий «понятность» перевода в наибольшем числе случаев. При этом подходе предпочтение отдавалось всегда тому варианту, который был по форме ближе всего к исходной английской конструкции: для того, чтобы избежать трансформаций при переводе, которые всегда приводят к появлению «шумов» и резкому увеличению вычислительных затрат и, соответственно, программистских усилий.

Актуальность проблемы моделирования трансформаций глагольных и именных конструкций для систем машинного перевода и извлечения знаний из текстов обусловлена тем, что до сих пор эти явления мало исследованы с точки зрения возможностей их компьютерных реализаций и, соответственно, недостаточно учтены в действующих системах

машинного перевода. Настоятельная потребность в создании функционально-семантических представлений глагольно-именных трансформаций также вызвана тем, что дальнейшее развитие систем машинного перевода ведется с использованием машинного обучения на параллельных корпусах и правила, задающие функциональную синонимию языковых конструкций позволяют извлечь необходимую информацию и избежать формирования избыточных правил и «шумов».

3. *Перевод фразеологических сочетаний в системах машинного перевода.*

Проблема машинного перевода **идиом** заключается в том, что не всегда удается дать точный перевод, руководствуясь обычными правилами. Однако следует принять во внимание, что идиомы должны быть выявлены на начальном этапе во избежание их утери, и обрабатываться они должны как одно слово.

Разрешение идиоматичности является одной из формальных операций, обеспечивающих анализ и синтез в системах машинного перевода, и производится либо с помощью стандартных грамматических и лексических программ анализа текста, работающих совместно с автоматическим словарем, либо путем прямого соотнесения входного и выходного сегментов. Во втором случае и входной, и выходной сегменты рассматриваются как неделимые обороты. Один или несколько выходных оборотов или словоформ, поставленных в соответствие каждому входному обороту, составляют автоматический словарь оборотов.

В результате анализа идиоматическим выражениям приписывается определенный цифровой эквивалент, и они исключаются из дальнейшего грамматического анализа.

Составлению алгоритма поиска и перевода оборотов в тексте предшествует лингвистическое исследование их дистрибуции. Следует выяснить посредством синтаксического анализа, является ли оборот цельным и включает ли он в себя изменяемые формы. Оборот считается цельным, если он имеет неизменный состав и между его элементами нельзя вставить другие единицы. Если оборот не является цельным, следует учитывать это при составлении алгоритма. Таким образом, для обработки оборотов, разорванных другими членами предложения, необходимы данные синтаксического анализа.

При создании систем машинного перевода текстов, содержащих идиоматические выражения, необходимо руководствоваться следующими принципами:

1. Основными единицами языка и речи, которые следует включать в машинный словарь, должны быть фразеологические единицы (в частности, идиоматические выражения). Отдельные слова также могут включаться в словарь, но они должны использоваться только в тех случаях, когда не удается осуществить перевод, опираясь только на фразеологические единицы.

2. Наряду с идиоматическими выражениями, состоящими из непрерывных последовательностей слов, в системах машинного перевода следует использовать и так называемые "речевые модели" – фразеологические единицы с "пустыми местами", которые могут заполняться различными словами и словосочетаниями, порождая осмысленные отрезки речи.

3. Реальные тексты, независимо от их принадлежности к той или иной тематической области, обычно бывают политематическими, если они имеют достаточно большой объем. И отличаются они друг от друга не столько словарным составом, сколько распределениями вероятностей появления в них различных слов из общенационального словарного фонда. Поэтому машинный словарь, предназначенный для перевода текстов даже только из одной тематической области, должен быть политематическим, а для перевода текстов из различных предметных областей - тем более.

4. Необходимы машинные словари большого объема. Такие словари должны создаваться на основе автоматизированной обработки двуязычных текстов, являющихся переводами друг друга, и в процессе функционирования систем перевода.

5. Наряду с основным политематическим словарем большого объема, в системах фразеологического машинного перевода целесообразно использовать также набор небольших по объему дополнительных тематических словарей. Дополнительные словари должны содержать только ту информацию, которая отсутствует в основном словаре (например, информацию о приоритетных переводных эквивалентах словосочетаний и слов для различных предметных областей, если эти эквиваленты не совпадают с приоритетными переводными эквивалентами основного словаря).

6. Наряду с переводом текстов в автоматическом режиме, в системах фразеологического машинного перевода целесообразно предусмотреть и

интерактивный режим их работы. В этом режиме пользователь должен иметь возможность вмешиваться в процесс перевода и настраивать дополнительные машинные словари на тематику переводимых текстов.

ЛЕКЦИЯ 6

Использование параллельных корпусов текстов в машинном переводе. Память переводов

1. *Виды и структура параллельных корпусов текстов.*
2. *Технология памяти переводов.*
3. *Преимущества и недостатки технологии.*

1. *Виды и структура параллельных корпусов текстов.*

Многоязычный корпус текстов представляет собой несколько аналогичных по структуре одноязычных корпусов текстов. Для параллельных корпусов выделяется ряд подтипов: тексты на языке А и их переводы на язык В; тексты на языках А и В и их переводы соответственно на языки В и А; только переводные тексты на языках А, В, С, Х, если оригинальные тексты были написаны на языке D. Кроме того, к параллельным корпусам можно отнести диахронические корпуса, которые составляются из текстов на более ранней форме языка и их переводов на современный язык, транскрипционные корпуса текстов, включающие тексты на литературном языке, прочитанные носителями разных его диалектов. В качестве подтипов можно выделить «шумные» параллельные корпуса; с пропусками в переводе, без точного соответствия между оригиналом и переводом), «зеркальные» параллельные корпуса, состоящие из текстов на языках А и В и переводов этих текстов соответственно на языки В и А.

Важность параллельных корпусов обусловлена тем, что они позволяют объективно установить, как переводчики на практике преодолевают трудности, и использовать эти данные для разработки соответствующих реальности моделей для начинающих переводчиков. Они также играют важную роль в исследовании переводческой нормы в специфических социокультурных и исторических контекстах. Это делает многоязычные корпуса текстов во многих отношениях привлекательными для переводчиков-практиков. Параллельные корпуса остаются незаменимым источником данных как для проведения исследований в области прикладной лингвистики (апробация систем автоматизированного перевода, заполнение систем переводческой памяти, разработка систем автоматического поиска переводных эквивалентов и т.п.), так и для

контрастивных и переводоведческих исследований (сравнение структуры исходного текста и перевода, определение степени информационных потерь при переводе, изучение различных переводческих стратегий и т. п.)

При составлении параллельных корпусов, в отличие от одноязычных и сопоставительных корпусов текстов, следует учитывать фактор межкультурных связей. Множество текстов исходного языка составляют лишь те тексты, которые были переведены на второй язык, и, если межкультурные связи полностью отсутствуют, получение параллельного корпуса невозможно. Чем слабее межнациональные и культурные связи, тем меньше переводов выполняется и тем более проблематично составление полноценного параллельного корпуса. Тексты на исходном языке, хотя и являются первичными, отбираются с учетом переводного языка. Структура субкорпуса исходного языка определяется наличием или отсутствием переводов на переводной язык, а также тем, какого рода тексты переводятся. При составлении параллельных корпусов могут использоваться разнообразные языковые ресурсы: специальные тексты, тексты СМИ, научные тексты, художественные тексты, т.е. параллельный корпус должен обладать свойством репрезентативности.

Структурная организация корпуса может быть самая разная, в зависимости от прагматических целей его создателя или пользователя:

– в виде традиционного текста со ссылкой на перевод;

– в табличной "зеркальной" форме, что более удобно для восприятия и сравнения;

– в виде базы данных (структура, применимая только при автоматической обработке).

Важным понятием является выравнивание текста. Выравнивание параллельного текста – это идентификация соответствующих друг другу предложений в обеих половинах параллельного текста. Выравнивание параллельного корпуса на уровне предложений является необходимой предпосылкой для различных аспектов лингвистических исследований. В процессе перевода предложения могут разделяться, сливаться, удаляться, вставляться или менять последовательность. В связи с этим выравнивание часто становится сложной задачей.

Параллельные корпуса текстов-образцов (в виде базы данных) особенно полезны в том случае, когда переводчик работает со строго нормированными (конвенциональными) текстами, жанрово-стилистическое и стилистическое оформление таких текстов практически не допускает варьирования, отступления от определенных

социокультурных норм. Это тексты деловой переписки, тексты-рецепты, тексты-прогнозы погоды, тексты-контракты и т.д. Тексты различных стилей различаются как словарем лексических единиц, употребляемых в определенных текстах, так и грамматическими и синтаксическими структурами предложений, заключенных в них. Параллельные корпуса текстов-образцов и их типологические модели-характеристики, составленные на этапе предпереводческого анализа исходного текста, могут служить для переводчика и студента таким же эффективным вспомогательным средством, как и различного рода словари.

2. Технология памяти переводов.

Идея параллельного корпуса имеет много общего с концепцией **памяти переводов**. Главное различие между ними в том, что память переводов представляет собой базу данных, в которой сегменты текста (соответствующие друг другу предложения) расположены таким способом, при котором они не связаны с оригинальным контекстом, то есть оригинальная последовательность предложений теряется. Параллельный корпус же сохраняет изначальную последовательность предложений.

Память переводов – база данных, содержащая набор ранее переведенных текстов. Одна запись в такой базе данных соответствует сегменту или «единице перевода», за которую обычно принимается одно предложение. Если единица перевода исходного текста в точности совпадает с единицей перевода, хранящейся в базе, она может быть автоматически подставлена в перевод. Новый сегмент может также слегка отличаться от хранящегося в базе. Такой сегмент может быть также подставлен в перевод, но переводчик будет должен внести необходимые изменения. Помимо ускорения процесса перевода повторяющихся фрагментов и изменений, внесенных в уже переведенные тексты (например, новых версий программных продуктов или изменений в законодательстве), системы памяти переводов также обеспечивают единообразие перевода терминологии в одинаковых фрагментах, что особенно важно при техническом переводе.

Основой функционирования любой системы памяти переводов являются ранее переведенные тексты. Множество этих текстов постоянно пополняется новыми переводами, вследствие чего процент автоматически переводимых сегментов постепенно растет. Это означает, что для наиболее эффективного использования памяти переводов все тексты должны содержать достаточное количество похожих фраз. Такое положение вещей

имеет место в документации на различного рода продукты. Это обусловлено двумя факторами. Во-первых, документацию принято составлять максимально простым языком, лаконично и в строгих терминах. Во-вторых, с появлением новых версий и модификаций поставляемого потребителям продукта содержание документации меняется лишь в незначительной степени. Память переводов в подобных случаях избавляет переводчика от необходимости по несколько раз переводить идентичные фрагменты текста, входящие в разные документы.

3. 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.

II ПРАКТИЧЕСКИЙ РАЗДЕЛ

2.1 Планы и темы практических занятий

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ №1

Машинный перевод как направление искусственного интеллекта

Вопросы:

1. Определения машинного перевода.
2. История разработок предметной области.
3. Предмет и объект научного направления.
4. Преимущества машинного перевода, цели разработки и сферы использования.

Задания:

1. Подготовить устные сообщения по машинному переводу первого, второго, третьего поколений.
2. Подобрать примеры использования машинного перевода представителями разных профессий, обосновать его необходимость.
3. Оценить перспективы машинного перевода, найти примеры полностью автоматического перевода в научно-фантастической литературе, обосновать реальность (нереальность) существования таких систем в будущем.

ПРАКТИЧЕСКИЕ ЗАНЯТИЯ № 2-3

Виды и стратегии машинного перевода

Вопросы:

1. Классификация переводов по различным критериям.
2. Типология перевода в зависимости от степени автоматизации: ручной, автоматизированный, автоматический перевод.
3. Стратегии машинного перевода: пословный перевод, перевод с использованием трансфера, с использованием интерлингвы, перевод, основанный на глубинном лингвистическом анализе, память переводов.

Задания:

1. Оценить возможности машинного перевода разных типов, оценить необходимость его использования.
2. Провести сравнительный анализ переводов текстов различных типов и жанров, сделать выводы.

3. Проанализировать работу систем машинного перевода, основанных на различных стратегиях, провести сравнительный анализ, сделать выводы.

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ № 4

Структура систем машинного перевода

Вопросы:

1. Структура лингвистической базы знаний систем машинного перевода.
2. Компоненты лингвистической базы данных.
3. Структура лингвистического процессора.

Задания:

1.Разработать структуру лингвистической базы знаний русско-английской системы машинного перевода.

2. Выявить отличия в структуре базы знаний для русско-английской и англо-русской систем машинного перевода.

3.Выявить особенности морфологического классификатора для английского языка в сравнении с русским.

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ № 5

Двуязычные словари в составе систем машинного перевода

Вопросы:

1. Особенности двуязычных машинных словарей в сравнении с одноязычными.
2. Представление лексических единиц в двуязычных машинных словарях.
3. Принципы отбора лексики для переводных словарей.
4. Кодирование лингвистической информации в словарях на морфологическом, синтаксическом, семантическом уровнях.

Задания:

1. Подобрать 200 русских лексем различных самостоятельных частей речи на основе частотных словарей.
2. Выделить все типы словоизменения для выбранных лексем.
3. Подобрать к ним переводные эквиваленты на английском языке с учетом многозначности.

4. Построить словарные статьи русско-английского машинного словаря по разным принципам (словарь словоформ, словарь квизиоснов).

5.Указать в словарях морфологическую, синтаксическую, семантическую информацию о лексемах.

ПРАКТИЧЕСКИЕ ЗАНЯТИЯ № 6-7

Алгоритм машинного перевода, основанного на лингвистическом анализе

Вопросы:

1. Этапы морфологического, синтаксического, семантического анализа исходного текста.

2. Синтез текста на переводном языке: семантическое представление текста, построение синтаксических структур, морфологический синтез.

Задания:

1. Перевести текст с помощью различных систем машинного перевода.

2. Выявить сложности, с которыми сталкиваются системы.

3. Провести морфологический анализ исходного текста, построить деревья зависимостей предложений, установить семантические связи внутри каждого из предложений.

4.Назвать этапы синтеза переводного текста, на которых выявленные ошибки могут быть исправлены.

5.Разработать рекомендации по улучшению качества работы систем машинного перевода.

Пример текста для анализа:

Thermohaline circulation produces great vertical currents' hat flow from the surface to the ocean bottom and back. The currents largely result from differences in water temperature and salinity. The currents move sluggishly from the polar regions, along the sea floor, and back to the surface. In the polar regions, the surface waters become colder and saltier. Being colder and saltier makes these waters heavier, and they gradually flow back toward the surface and replace the surface waters that sink. For example, as a warm ocean current and a cold ocean current meet together, the warm water will always follow cold water and moving around the ocean. The most important is the plankton does not like to stay in the cold place. They will follow the warm current. As the result the small fish like shrimp that eats plankton, will follow the plankton. Similarly

the bigger fish that eats small fish and shrimp will just follow also and these fishes became a chain.

ПРАКТИЧЕСКИЕ ЗАНЯТИЯ № 8-9

Лингвистические проблемы машинного перевода

Вопросы:

1. Проблемы на морфологическом и синтаксическом уровнях.
2. Формальные и содержательные несоответствия близкородственных и неблизкородственных языков.
3. Проблемы машинного перевода на семантическом и прагматическом уровнях.
4. Учет контекста и экстралингвистических факторов при разрешении полисемии.
5. Варианты решения семантических сложностей и повышения качества машинного перевода.

Задания:

1. Выбрать 5 русских и 5 английских отрывков текстов разных стилей: научно-технического, публицистического, художественного, разговорного, стихотворного.
2. Перевести с помощью пяти различных систем машинного перевода.
3. Выявить ошибки в переводах на различных языковых уровнях, предложить пути их решения.
4. Найти в исходных текстах омонимы (слова, имеющие несколько переводных эквивалентов), составить для них статьи контекстологического словаря.

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ № 10

Память переводов

Вопросы:

1. Системы машинного перевода, основанные на примерах.
2. Принципы построения параллельного корпуса текстов для систем памяти переводов.
3. Структура систем, использующих параллельные корпусы текстов и алгоритм работы.
4. Сфера применения памяти переводов.

Задания:

1.Разработать алгоритм построения параллельного двуязычного корпуса текстов.

2. Подобрать текст на русском языке и его литературный перевод на английский, провести выравнивание текстов по предложениям.

3.Найти онлайн-систему памяти переводов, проанализировать ее работу по различным критериям.

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ № 11

Оценка качества машинного перевода

Вопросы:

1. Критерии оценки качества машинного перевода.
2. Шкала оценки качества перевода.
3. Качество перевода на различных языковых уровнях.

Задания:

1. Познакомиться со статьей Л.Маркеса «Automatic Evaluation of Machine Translation Quality» [http://www.dialog-21.ru/digests/dialog2013/materials/pdf/1_MarquezL.pdf].

2. Используя сервис Asiya [http://asiya.lsi.upc.edu/demo/], оценить качество перевода нескольких текстов, подобранных самостоятельно.

3. Описать сервис Asiya, указать его пользу, новаторство, преимущества и недостатки.

4. Разработать анкету для испытуемых по определенному тексту, используя методику «изучение ответов по оригиналу текста человека, читавшего только перевод», провести опрос, сделать выводы.

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ № 12

Связь машинного перевода с другими областями знаний

Вопросы:

1. Машинный перевод в контексте исследований в области искусственного интеллекта.

2. Машинный перевод и компьютерная лингвистика.

3. Связь машинного перевода с математической лингвистикой и психолингвистикой.

Задания:

1.Провести исследование научной литературы по данной тематике, подобрать примеры и подготовить реферат.

2. Подготовить выступление перед аудиторией, провести обсуждение докладов.

III РАЗДЕЛ КОНТРОЛЯ ЗНАНИЙ

3.1 Примерный перечень вопросов для подготовки к экзамену

1. Машинный перевод как научное направление, цели и задачи. Предмет и объект данной области знаний.
2. История машинного перевода. Три поколения развития.
3. Преимущества машинного перевода и цели разработки.
4. Виды, формы, типы ручного перевода. Типология перевода в зависимости от степени автоматизации: ручной, автоматизированный, автоматический перевод.
5. Стратегии машинного перевода: пословный перевод, перевод с использованием трансфера, с использованием интерлингвы, перевод, основанный на глубинном лингвистическом анализе, память переводов.
6. Общий алгоритм решения задачи машинного перевода.
7. Синтез текста на переводном языке: семантическое представление текста, построение синтаксических структур, морфологический синтез.
8. Структура лингвистической базы знаний систем машинного перевода: лингвистическая база данных и лингвистический процессор.
9. Компоненты лингвистической базы данных систем машинного перевода.
10. Морфологический словарь в структуре лингвистической базы данных.
11. Синтаксический словарь в структуре лингвистической базы данных.
12. Семантический словарь в структуре лингвистической базы данных.
13. Структура лингвистического процессора систем машинного перевода.
14. Морфологический и синтаксический анализ исходного текста в системах машинного перевода.
15. Семантический и прагматический анализ исходного текста в системах машинного перевода.
16. Лингвистические проблемы машинного перевода на морфологическом уровне. Пути решения проблем.
17. Лингвистические проблемы машинного перевода на синтаксическом уровне. Пути решения проблем.
18. Лингвистические проблемы машинного перевода на семантическом уровне. Пути решения проблем.
19. Лингвистические проблемы машинного перевода на прагматическом уровне. Пути решения проблем.
20. Синтаксические трансформации в машинном переводе.
21. Перевод фразеологических сочетаний в машинном переводе.

22. Учет контекста и экстралингвистических факторов при разрешении полисемии.
23. Системы машинного перевода, основанные на примерах. Принципы построения корпуса текстов для систем памяти переводов.
24. Структура систем и алгоритм работы памяти переводов. Сфера применения.
25. Контекстологический словарь в структуре систем машинного перевода.
26. Критерии оценки качества машинного перевода. Оценка перевода испытуемыми. Шкала оценки качества перевода.
27. Критерии оценки качества машинного перевода. Статистический критерий. Изучение ответов по оригиналу текста человека, читавшего только перевод.
28. Оценка качества по трудозатратам на корректировку машинного перевода. Качество перевода на различных языковых уровнях.
29. Машинный перевод в контексте исследований в области искусственного интеллекта. Машинный перевод и компьютерная лингвистика.
30. Машинный перевод и информационный поиск.
31. Связь машинного перевода с математической лингвистикой и психолингвистикой.
32. Машинный перевод и теория перевода.
33. Машинный перевод в деятельности переводчика.

3.2 Образцы заданий

Пример контрольной работы по теме практических занятий №8-9

1 . Проанализируйте текст, выявите лингвистические проблемы, с которыми может столкнуться система автоматического перевода, предложите пути их решения.

Студенты разработали виртуального гида — специальное мобильное приложение, благодаря которому посетители музеев не будут нуждаться в экскурсоводе, а со смартфона можно получить всю необходимую информацию обо всех экспонатах. Родители с помощью сотового телефона заглядывают в дневник ребенка, началось внедрение электронных рецептов в аптеках. Глава государства в Послании к белорусскому народу и Национальному собранию поставил задачу

повсеместного внедрения новых информационных технологий. А что агропромышленный комплекс, далеко ли продвинулся в этой сфере?

У многих агрономов или зоотехников теперь есть смартфон. И, надеюсь, компьютер на рабочем столе. Но для чего они используются? Вы видели когда-нибудь, чтобы специалист использовал их для профессиональных целей, например, подбора вакцины животным или определения плодородия конкретного поля? Лично я нет. Хотя знаком со многими успешными агроменеджерами. Но никого из них нельзя винить в том, что отстали от жизни. Специалистам просто не предлагают соответствующих мобильных и иных информационных программ. А заняться этим давно пора и главному аграрному ведомству, и различным научным учреждениям, которых у нас немало.

Обратимся к опыту других стран, на которые мы хотим равняться в развитии АПК. Например, в США множество программ (софтов) от министерства сельского хозяйства, государственных и частных исследовательских центров позволяют оперативно получать любую информацию.

В Германии существует бесплатное приложение для отслеживания отелов. По сути, это дневник беременности коровы, который можно вести в любом виде компьютера. Он рассчитывает примерную дату отела и вовремя выдает напоминание об этом.

Если мы не преуспели в этом сегменте информационных технологий, так может, есть прогресс в других? Увы! Государственная программа развития аграрного бизнеса на эту пятилетку предусматривает переход к точному земледелию. Появились навигационные системы, которые позволяют машинам для подкормки растений и внесения гербицидов не оставлять пустых участков или не обрабатывать их дважды. Вот, пожалуй, и все.

Между тем суть точного земледелия в ином. Питательные вещества на отдельных участках поля распределены неравномерно. Однако удобрения вносятся в усредненной дозе. Поэтому одним растениям их не хватает, у других — излишек. Это негативно отражается на урожае и приводит к перерасходу туков. Для решения проблемы нужно создать электронную карту плодородия поля, а затем на каждый клочок с помощью спутниковой навигации вносить удобрений столько, сколько требуется.

Такой подход активно используется во многих странах. Например, в Германии более половины сельхозпроизводителей применяют систему

точного земледелия. У нас же дальше экспериментов дело не пошло. Ученые предложили несколько вариантов электронного мониторинга почвы и урожайности, однако ни одно хозяйство серьезно за них не взялось.

Порой приходится слышать: какие там IT-технологии, не до жиру, быть бы живу. Однако, думается, одна из проблем нашего АПК в неумении считать деньги. Эксперты доказали: внедрение элементов точного земледелия дает прибавку урожая на 30 процентов, экономит столько же минеральных удобрений, наполовину снижает расход гербицидов. Это означает, что небольшие затраты на инновации быстро себя окупают. Возможно, и господдержку стоит не просто перечислять в хозяйства деньгами, а направлять хотя бы частично на внедрение точного земледелия. Как говорят, давать удочку, а не рыбу. Добавлю: и учить ловить крупную. (https://www.sb.by/articles/informtekhnologii-v-apk-pochti-ne-dayut-vskhodov.html)

2. Проанализируйте перевод, сделанный автоматическим переводчиком. Найдите ошибки и укажите, с помощью каких механизмов их можно было бы избежать.

Оригинал: *A man walks into a doctor's office. He has a cucumber up his nose, a carrot in his left ear and a banana in his right ear.*

"What's the matter with me?" he asks the doctor.

The doctor replies, "You're not eating properly."

Перевод*: Человек заходит в кабинет врача. У него есть огурцы нос, морковь в его левое ухо и банан в правом ухе. "Что это со мной?" Он спрашивает врач. Врач отвечает: «Ты не есть правильно".*

Оригинал: *One day an Englishman, a Scotsman, and an Irishman walked into a pub together. They each bought a pint of Guinness. Just as they were about to enjoy their creamy beverage, three flies landed in each of their pints, and were stuck in the thick head. The Englishman pushed his beer away in disgust. The Scotsman fished the fly out of his beer, and continued drinking it, as if nothing had happened. The Irishman, too, picked the fly out of his drink, held it out over the beer, and started yelling, "SPIT IT OUT, SPIT IT OUT YOU ...."*.

Перевод*: Однажды англичанин, шотландец, ирландец и вошел в паб вместе. У каждого из них купили пинту Guinness. Так же, как они собираются, чтобы насладиться их сливочный напиток, три мухи приземлились в каждом из пунктов, и застрял в гуще голову. Англичанин толкнул его пивом с отвращением. The Scotsman ловил муху из своего пива, и продолжал пить ее, как будто ничего не случилось. Ирландец тоже взял*

муху из своего напитка, протянул ее за пивом, и начал кричать: "выплюнуть, выплюнуть ВЫ ....".

3.3 Критерии оценки знаний и компетенции студентов

10 баллов - десять:

- систематизированные, глубокие и полные знания по всем разделам учебной программы, а также по основным вопросам, выходящим за ее пределы;
- точное использование научной терминологии (в том числе на иностранном языке), стилистически грамотное, логически правильное изложение ответа на вопросы;
- безупречное владение инструментарием учебной дисциплины, умение его эффективно использовать в постановке и решении научных и профессиональных задач;
- выраженная способность самостоятельно и творчески решать сложные проблемы в нестандартной ситуации;
- полное и глубокое усвоение основной и дополнительной литературы, рекомендованной учебной программой дисциплины;
- умение ориентироваться в теориях, концепциях и направлениях по изучаемой дисциплине и давать им критическую оценку, использовать научные достижения других дисциплин;
- творческая самостоятельная работа на практических, лабораторных занятиях, активное участие в групповых обсуждениях, высокий уровень культуры исполнения задании.

9 баллов - девять:

- систематизированные, глубокие и полные знания по всем разделам учебной программы;
- точное использование научной терминологии (в том числе на иностранном языке), стилистически грамотное, логически правильное изложение ответа на вопросы;

- владение инструментарием учебной дисциплины, умение его эффективно использовать в постановке и решении научных и профессиональных задач;

- способность самостоятельно и творчески решать сложные проблемы в нестандартной ситуации в рамках учебной программы;
- полное усвоение основной и дополнительной литературы, рекомендованной учебной программой дисциплины;
- умение ориентироваться в основных теориях, концепциях и направлениях по изучаемой дисциплине и давать им критическую оценку;
- самостоятельная работа на практических, лабораторных занятиях, творческое участие в групповых обсуждениях, высокий уровень культуры исполнения заданий.

8 баллов - восемь:

- систематизированные, глубокие и полные знания по всем поставленным вопросам в объеме учебной программы;
- использование научной терминологии, стилистически грамотное, логически правильное изложение ответа на вопросы, умение делать обоснованные выводы;
- владение инструментарием учебной дисциплины (методами комплексного анализа, техникой информационных технологий), умение его использовать в постановке и решении научных и профессиональных задач;
- способность самостоятельно решать сложные проблемы в рамках учебной программы;
- усвоение основной и дополнительной литературы, рекомендованной учебной программой дисциплины;
- умение ориентироваться в основных теориях, концепциях и направлениях по изучаемой дисциплине и давать им критическую оценку с позиций государственной идеологии (по дисциплинам социально-гуманитарного цикла);
- активная самостоятельная работа на практических, лабораторных занятиях, систематическое участие в групповых обсуждениях, высокий уровень культуры исполнения заданий.

7 баллов - семь:

- систематизированные, глубокие и полные знания по всем разделам учебной программы;

- использование научной терминологии (в том числе на иностранном языке), лингвистически и логически правильное изложение ответа на вопросы, умение делать обоснованные выводы;
- владение инструментарием учебной дисциплины, умение его использовать в постановке и решении научных и профессиональных задач;
- усвоение основной и дополнительной литературы, рекомендованной учебной программой дисциплины;
- умение ориентироваться в основных теориях, концепциях и направлениях по изучаемой дисциплине и давать им критическую оценку;
- самостоятельная работа на практических, лабораторных занятиях, участие в групповых обсуждениях, высокий уровень культуры исполнения заданий.

6 баллов - шесть:

- достаточно полные и систематизированные знания в объеме учебной программы;
- использование необходимой научной терминологии, стилистически грамотное, логически правильное изложение ответа на вопросы, умение делать обоснованные выводы;
- владение инструментарием учебной дисциплины, умение его использовать в решении учебных и профессиональных задач;
- способность самостоятельно применять типовые решения в рамках учебной программы;
- усвоение основной литературы, рекомендованной учебной программой дисциплины;
- умение ориентироваться в базовых теориях, концепциях и направлениях по изучаемой дисциплине и давать им сравнительную оценку;
- активная самостоятельная работа на практических, лабораторных занятиях, периодическое участие в групповых обсуждениях, высокий уровень культуры исполнения заданий.

5 баллов - пять:

- достаточные знания в объеме учебной программы;
- использование научной терминологии, стилистически грамотное, логически правильное изложение ответа на вопросы, умение делать выводы;

- владение инструментарием учебной дисциплины, умение его использовать в решении учебных и профессиональных задач;
- способность самостоятельно применять типовые решения в рамках учебной программы;
- усвоение основной литературы, рекомендованной учебной программой дисциплины;
- умение ориентироваться в базовых теориях, концепциях и направлениях по изучаемой дисциплине и давать им сравнительную оценку;
- самостоятельная работа на практических, лабораторных занятиях, участие в групповых обсуждениях, высокий уровень культуры исполнения заданий.

4 балла - четыре, УДОВЛЕТВОРИТЕЛЬНО:

- достаточный объем знаний в рамках образовательного стандарта;
- усвоение основной литературы, рекомендованной учебной программой дисциплины;
- использование научной терминологии, стилистическое и логическое изложение ответа на вопросы, умение делать выводы без существенных ошибок;
- владение инструментарием учебной дисциплины, умение его использовать в решении стандартных (типовых) задач;
- умение под руководством преподавателя решать стандартные (типовые) задачи;
- умение ориентироваться в основных теориях, концепциях и направлениях по изучаемой дисциплине и давать им оценку;
- работа под руководством преподавателя на практических и лабораторных занятиях, допустимый уровень культуры исполнения заданий.

3 балла - три, НЕУДОВЛЕТВОРИТЕЛЬНО:

- недостаточно полный объем знаний в рамках образовательного стандарта;
- знание части основной литературы, рекомендованной учебной программой дисциплины;
- использование научной терминологии, изложение ответа па вопросы с существенными лингвистическими и логическими ошибками;

- слабое владение инструментарием учебной дисциплины, некомпетентность в решении стандартных (типовых) задач;
- неумение ориентироваться в основных теориях, концепциях и направлениях изучаемой дисциплины;
- пассивность на практических и лабораторных занятиях, низкий уровень культуры исполнения заданий.

2 балла - два, НЕУДОВЛЕТВОРИТЕЛЬНО:

- фрагментарные знания в рамках образовательного стандарта;
- знания отдельных литературных источников, рекомендованных учебной программой дисциплины;
- неумение использовать научную терминологию дисциплины, наличие в ответе грубых стилистических и логических ошибок;
- пассивность на практических и лабораторных занятиях, низкий уровень культуры исполнения заданий.

1 балл - один, НЕУДОВЛЕТВОРИТЕЛЬНО:

- отсутствие знаний и компетенций в рамках образовательного стандарта или отказ от ответа.

IV ВСПОМОГАТЕЛЬНЫЙ РАЗДЕЛ

4.1 Учебная программа дисциплины

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Основная цель курса «машинный перевод» состоит в формировании у студентов знаний о принципах и методах разработки систем машинного перевода, об их лингвистическом обеспечении, об истории разработок и существующих системах, а также в формировании умений и навыков формализации знаний о языке и алгоритмизацией лингвистического анализа с целью создания лингвистического процессора систем машинного перевода.

Задачи курса:

обучить студентов основным принципам создания и использования систем машинного перевода;

ознакомить студентов с теоретическими основами, методами и средствами формализации языка;

научить студентов составлению лингвистических алгоритмов анализа слова и его значения, структуры и семантики предложения и текста при создании лингвистического процессора системы машинного перевода.

Курс «Машинный перевод» тесно связан с изучаемыми на отделении СИЯ дисциплинами «Компьютерная лингвистика», «Автоматическая обработка естественного языка», «Проблемы искусственного интеллекта», «Квантитативная лингвистика», «Инженерия знаний»; комплекс названных дисциплин призван сформировать целостное представление о современных исследованиях в области искусственного интеллекта.

Курс предполагает углубленное знакомство с современными направлениями использования языка в информационных технологиях; с формализацией знаний о языке и алгоритмизацией лингвистического анализа; с принципами создания лингвистических банков данных и баз знаний. Приложение знаний специализации может найти применение в преподавании иностранного языка с использованием новых информационных технологий в школе и вузе; в создании систем машинного перевода; в создании лингвистических банков данных и научной работе на базе этих банков данных.

Курс «Машинный перевод» состоит из цикла лекций и цикла практических занятий. Лекции носят проблемный характер, связаны с наиболее важными направлениями исследований и призваны предоставить

студентам знания в предметной области. Практические занятия направлены на формирование умений и навыков создания лингвистических банков данных, лингвистического процессора систем машинного перевода, формализации знаний о естественном языке.

Курс включает:
Лекции **16** (часов)
Практические (семинарские) занятия **12** (часов)
Контролируемая самостоятельная работа **6** (часов)

ВСЕГО ЧАСОВ: **72 часа**

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

№ пп	Наименование разделов, тем	Количество часов				
		Аудиторные				
		Лекции	Практич./ семинары	Лабор. занят.	КСР	Самост. работа
1	Машинный перевод как направление искусственного интеллекта	2				2
2	Виды и стратегии машинного	2	2			2
3	Алгоритм машинного перевода, основанного на лингвистическом анализе	2	2			2
4	Структура систем машинного	2	2		2	2
5	Лингвистические проблемы машинного перевода	4	4		2	2
6	Память переводов	2	2			2
7	Связь машинного перевода с другими областями знаний	2			2	2
	Итого:	16	12		6	16

Раздел 1. МАШИННЫЙ ПЕРЕВОД КАК НАПРАВЛЕНИЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Машинный перевод как научное направление, цели и задачи. Предмет и объект данной области знаний. История машинного перевода. Преимущества машинного перевода и цели разработки.

Раздел 2. ВИДЫ И СТРАТЕГИИ МАШИННОГО ПЕРЕВОДА

Виды, формы, типы ручного перевода. Типология перевода в зависимости от степени автоматизации: ручной, автоматизированный, автоматический перевод. Стратегии машинного перевода: пословный перевод, перевод с использованием трансфера, с использованием интерлингвы, перевод, основанный на глубинном лингвистическом анализе, память переводов.

Раздел 3. АЛГОРИТМ МАШИННОГО ПЕРЕВОДА, ОСНОВАННОГО НА ЛИНГВИСТИЧЕСКОМ АНАЛИЗЕ

Этапы морфологического, синтаксического, семантического анализа исходного текста. Синтез текста на переводном языке: семантическое представление текста, построение синтаксических структур, морфологический синтез.

Раздел 4. СТРУКТУРА СИСТЕМ МАШИННОГО ПЕРЕВОДА

Структура лингвистической базы знаний систем машинного перевода: лингвистическая база данных и лингвистический процессор. Компоненты лингвистической базы данных. Структура лингвистического процессора.

Раздел 5. ЛИНГВИСТИЧЕСКИЕ ПРОБЛЕМЫ МАШИННОГО ПЕРЕВОДА

Проблемы на морфологическом и синтаксическом уровнях. Формальные и содержательные несоответствия близкородственных и неблизкородственных языков. Пути решения проблем. Проблемы машинного перевода на семантическом и прагматическом уровнях. Учет контекста и экстралингвистических факторов при разрешении полисемии. Варианты решения семантических сложностей и повышения качества машинного перевода.

Раздел 6. ПАМЯТЬ ПЕРЕВОДОВ

Системы машинного перевода, основанные на примерах. Принципы построения корпуса текстов для систем памяти переводов. Структура систем и алгоритм работы. Сфера применения памяти переводов.

Раздел 7. СВЯЗЬ МАШИННОГО ПЕРЕВОДА С ДРУГИМИ ОБЛАСТЯМИ ЗНАНИЙ

Машинный перевод в контексте исследований в области искусственного интеллекта. Машинный перевод и компьютерная лингвистика. Связь машинного перевода с математической лингвистикой и психолингвистикой.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА

Номер раздела, темы, занятия	Название раздела, темы, занятия; перечень изучаемых вопросов	Количество аудиторных часов				Материальное обеспечение занятия (наглядные, методические пособия и ДР-)	Литература	Форма контроля знаний
		лекции	практические (семинарские) занятия	лабораторные занятия	управляемая самостоятельная работа студента			
1	2	3	4	5	6	7	8	9
1.	**Машинный перевод как направление искусственного интеллекта**							
	Машинный перевод как научное направление, цели и задачи. Предмет и объект данной области знаний. История машинного перевода. Преимущества машинного перевода и цели разработки.	2				Электронный конспект лекций Презентации лекций	1,2,3,4; 7(доп.)	Устный опрос
2	**Виды и стратегии машинного перевода**							
	Виды, формы, типы ручного перевода. Типология перевода в зависимости от степени автоматизации: ручной, автоматизированный, автоматический перевод. Стратегии машинного перевода: пословный перевод, перевод с использованием трансфера, с использованием интерлингвы, перевод, основанный на глубинном лингвистическом анализе, память переводов.	2	2			Электронный конспект лекций Презентации лекций	1,2,4,5; 6(доп.)	Контр. работа
3	**Алгоритм машинного перевода, основанного на лингвистическом анализе**							
	Этапы морфологического, синтаксического, семантического анализа исходного текста. Синтез текста на переводном языке: семантическое представление текста, построение синтаксических структур, морфологический синтез.	2	2			Электронный конспект лекций Презентации лекций	3,5,6,7; 2(доп.)	Устный опрос

4	Структура систем машинного перевода							
	Структура лингвистической базы знаний систем машинного перевода: лингвистическая база данных и лингвистический процессор. Компоненты лингвистической базы данных. Структура лингвистического процессора.	2	2		2	Электронный конспект лекций Презентации лекций	4,5,6,7; 5(доп.)	Тест
5	Лингвистические проблемы машинного перевода							
5.1	Проблемы на морфологическом и синтаксическом уровнях. Формальные и содержательные несоответствия близкородственных и неблизкородственных языков. Пути решения проблем.	2	2			Электронный конспект лекций Презентации лекций	1,3,6,7; 3(доп.)	Устный опрос
5.2	Проблемы машинного перевода на семантическом и прагматическом уровнях. Учет контекста и экстралингвистических факторов при разрешении полисемии. Варианты решения семантических сложностей и повышения качества машинного перевода.	2	2		2	Электронный конспект лекций Презентации лекций	1,2,3,5	Контр. работа
6	Память переводов							

	Системы машинного перевода, основанные на примерах. Принципы построения корпуса текстов для систем памяти переводов. Структура систем и алгоритм работы. Сфера применения памяти переводов.	2	2			Электронный конспект лекций Презентации лекций	3,6;3, 4(доп.)	Устный опрос
7	Связь машинного перевода с другими областями знаний							
	Машинный перевод в контексте исследований в области искусственного интеллекта. Машинный перевод и компьютерная лингвистика. Связь машинного перевода с математической лингвистикой и психолингвистикой.	2			2	Электронный конспект лекций Презентации лекций	1,2,5; 1(доп.)	Устный опрос
	ВСЕГО:	16	12		6			

ЛИТЕРАТУРА

Перечень основной литературы:

1. Семенов А.Л. Современные информационные технологии и перевод. – М., 2008.
2. Зубов А.В., Зубова И.И. Основы искусственного интеллекта для лингвистов. – М., 2007.
3. Мамедова М.Т. Машинный перевод. Эволюция и основные аспекты моделирования. – Баку, 2006
4. Нелюбин Л.Л. Компьютерная лингвистика и машинный перевод. – М., 1991.
5. Баранов А.К. Введение в прикладную лингвистику. – М., 2001.
6. Марчук Ю.Н. Проблемы машинного перевода. – М., 1983
7. Кулагина О. С. Исследования по машинному переводу. – М., 1979.

Перечень дополнительной литературы:

1. Машинный перевод и прикладная лингвистика. – М., 1980.
2. Рябцева Н.К. Информационные процессы и машинный перевод. – М., 1986.
3. Марчук Ю.Н. Основы компьютерной лингвистики. – М., 2000.
4. Степанова Д.В. Лингвистические аспекты перевода на русский язык английских терминологических словосочетаний с использованием корпуса параллельных текстов. – Мн., 2007.
5. Шаляпина З.М. Текст как объект автоматического перевода. – М., 1988
6. Леонтьева Н.Н., Шаляпина З.М. Современное состояние машинного перевода // Искусственный интеллект. Справочник. Кн.1. Системы общения и экспертные системы. – М., 1990.
7. Апресян Ю.Д. Лингвистический процессор для сложных информационных систем. – М.: Наука, 1992.

4.2 Библиография

1. Апресян Ю.Д. Лингвистический процессор для сложных информационных систем. – М.: Наука, 1992.
2. Баранов А.Н. Введение в прикладную лингвистику. М., 2001.
3. Белоногов Г. Г. Системы фразеологического машинного перевода политематических текстов [Электронный ресурс]. Режим доступа: http://www.a-z.ru/person/belonogov/index.htm
4. Борисевич А.Д. Англо-русский автоматический словарь оборотов: (К проблеме идиоматичности при обращении текста в системе «человек-машина-человек»): автореф. дис. на соиск. учен. степ. канд. филол. наук: (10663)/ А.Д. Борисевич – Минск: БГУ им. В.И.Ленина, 1972.
5. Зубов А.В., Зубова И.И. Основы искусственного интеллекта для лингвистов. М., 2007.
6. Козеренко Е.Б. Глагольно-именные трансформации при англо-русском машинном переводе [Электронный ресурс]. Режим доступа: http://www.dialog-21.ru/digests/dialog2007/materials/html/43.htm
7. Кулагина О. С. Исследования по машинному переводу. – М., 1979.
8. Леонтьева Н.Н., Шаляпина З.М. Современное состояние машинного перевода // Искусственный интеллект. Справочник. Кн.1. Системы общения и экспертные системы. - М., 1990.
9. Мамедова М.Т. Машинный перевод. Эволюция и основные аспекты моделирования – Баку, 2006
10. Марчук Ю.Н. Основы компьютерной лингвистики. М., 2000.
11. Марчук Ю.Н. Проблемы машинного перевода. М., 1983
12. Машинный перевод и прикладная лингвистика. М., 1980.
13. Нелюбин Л.Л. Компьютерная лингвистика и машинный перевод. М., 1991.
14. Панич Ю. В. Предварительная идентификация неоднозначного исходного текста и его перевод на другие языки с использованием системы согласованных словарей [Электронный ресурс]. Режим доступа: http://www.sciteclibrary.ru/rus/catalog/pages/9402.html
15. Рябцева Н.К. Информационные процессы и машинный перевод. М., 1986.
16. Семенов А.Л. Современные информационные технологии и перевод. М., 2008.

17. Степанова Д.В. Лингвистические аспекты перевода на русский язык английских терминологических словосочетаний с использованием корпуса параллельных текстов. Мн., 2007.
18. Шаляпина З.М. Текст как объект автоматического перевода. М., 1988.