МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Российский государственный гуманитарный университет»
(ФГБОУ ВО «РГГУ»)

ИНСТИТУТ ЛИНГВИСТИКИ
Учебно-научный центр компьютерной лингвистики

Введение в компьютерную лингвистику

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

45.04.03 Фундаментальная и прикладная лингвистика
Код и наименование направления подготовки/специальности

Фундаментальная и компьютерная лингвистика
Наименование направленности (профиля)/ специализации

Уровень высшего образования: магистратура

Форма обучения: очная

РПД адаптирована для лиц
с ограниченными возможностями
здоровья и инвалидов

Москва 2023

Введение в компьютерную лингвистику
Рабочая программа дисциплины

Составитель(и):

К.ф.н., доцент А.Ч.Пиперски

Ответственный редактор:

К.ф.н, доцент Н.А.Коротаев

УТВЕРЖДЕНО
Протокол заседания УНЦ компьютерной лингвистики
№ 6 от 12 апреля 2023 г.

ОГЛАВЛЕНИЕ

1. Пояснительная записка

1.1. Цель и задачи дисциплины

Целью дисциплины является ознакомление с методами и лингвистическими технологиями, применяемыми при создании компьютерных систем обработки текстов в научно-практической области исследований «компьютерная лингвистика», и рассматриваемых в сопоставлении с лингвистическими и культурно-лингвистическими свойствами языковых произведений – предложений и текстов, а также в связи с задачами обработки текстов как социальными запросами общества. Подача материала частично увязана с историей компьютерной лингвистики, что позволяет лучше представить качественную составляющую процессов, моделируемых современными лингвистическими технологиями, изучаемыми в других курсах по профилю «Компьютерная лингвистика», основанных преимущественно на эмпирических, в частности, статистических методах.

Курс направлен на решение следующих задач:

- создать представление о компьютерной лингвистике как новейшей научно-практической области исследований, ее возникновении в контексте смежных наук и ее современной организации;
- познакомить магистрантов с основными лингвистическими технологиями, реализующими анализ предложения (текста) по уровням лингвистической разметки и основными приемами автоматической генерации текстов;
- познакомить магистрантов с основными типами ресурсов, создающимися и использующимися компьютерными программами для решения конкретных задач в исследовательских целях, при разработке лингвистических технологий и в приложениях;
- соединить интуитивные и традиционные представления о свойствах естественно-языковых текстов со способами их формализации и моделирования в работах по компьютерной лингвистике;
- выработать у магистрантов элементарные практические навыки по применению компьютерно-лингвистических методов к языковому материалу и использованию лингвистических технологий.

1.2. Перечень планируемых результатов обучения по дисциплине, соотнесенных с индикаторами достижения компетенций

Компетенция	Индикаторы компетенций	Результаты обучения
ОПК-3 Способен выбирать оптимальные подходы и методы решения конкретных научных и прикладных задач в области лингвистики и информационных технологий	ОПК-3.1 Знает основные типы систем, использующих модули лингвистического анализа; основные принципы и методы компьютерного моделирования лингвистических задач	Знать: – основные типы систем, использующих модули лингвистического анализа; – основные принципы и методы компьютерного моделирования лингвистических задач; Уметь: – анализировать и сопоставлять различные типы систем, использующих модули лингвистического анализа; – использовать программы для автоматического анализа и обработки естественного языка;

		Владеть: – методами компьютерного моделирования лингвистических задач.
	ОПК-3.2 Умеет анализировать работу различных систем обработки текста и звучащей речи для выявления основных лингвистических компонентов и основных типов обработки текста, используемых в данных системах; подбирать необходимые лингвистические ресурсы для различных задач лингвистического обеспечения систем (например, лексикографических, задач морфологического анализа и т.п.)	Знать: – основные типы систем, использующих модули анализа и обработки текста и звучащей речи; – основные лингвистические компоненты и основные типы обработки текста; Уметь: – анализировать и сопоставлять различные типы систем обработки текста и звучащей речи; – подбирать необходимые лингвистические ресурсы для различных задач лингвистического обеспечения систем; Владеть: – навыками использования различных систем обработки текста и звучащей речи.

1.3. Место дисциплины в структуре образовательной программы

Дисциплина «Введение в компьютерную лингвистику» относится к обязательной части блока дисциплин учебного плана.

В результате освоения дисциплины формируются знания, умения и владения, необходимые для изучения следующих дисциплин и прохождения практик: Русская корпусная грамматика, Статистические методы в лингвистике, Автоматическая оценка сложности текстов, Автоматический семантический анализ.

2. Структура дисциплины

Общая трудоёмкость дисциплины составляет 3 з.е., 108 академических часов.

Структура дисциплины для очной формы обучения

Объем дисциплины в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении учебных занятий:

Семестр	Тип учебных занятий	Количество часов
1	Лекции	6
	Практические занятия	24
Всего:		30

Объем дисциплины (модуля) в форме самостоятельной работы обучающихся составляет 78 академических часов.

3. Содержание дисциплины

№	Наименование раздела дисциплины	Содержание
Часть 1. Задачи и методы обработки языковых произведений – анализ предложения		
1.	Компьютерная лингвистика (КЛ) в системе смежных наук. Краткая история и предмет КЛ.	Возникновение компьютерной лингвистики как новейшего научно-практического направления исследований, ее объект и предмет. КЛ и формы передачи информации. КЛ в системе смежных наук, общая структура и основные разделы КЛ. Теоретическая и прикладная КЛ.
2.	Уровни языка в традиционной лингвистике и уровневые модели естественного языка (ЕЯ) в КЛ. Задачи анализа предложения. Проблема неоднозначности.	Уровни языка в традиционной лингвистике и уровневые модели естественного языка в КЛ: уровневая модель «Смысл-Текст» И.А. Мельчука как структурная модель ЕЯ. Понятие лингвистического представления. «Системно-функциональная грамматика» М.А.К. Хэллидея как функциональная модель ЕЯ. Виды лингвистической информации, выражаемой в предложении: морфологическая, лексическая, синтаксическая, дискурсная, прагматическая и ее выявление методами КЛ. Сегментация текста и морфологический анализ. Конечные автоматы и регулярные выражения. Задачи синтаксического компонента и синтаксические представления. Задачи дискурсивного анализа. Многозначность элементов ЕЯ и проблема неоднозначности в КЛ, виды неоднозначности.
3.	«Понимание» текстов в узкой предметной области и неоднозначность. Метод шаблонов. Семантически ориентированный метод анализа.	Моделирование «понимания» в узкой предметной области как способ избежать решения проблемы неоднозначности. Составляющие «понимания» текста компьютером на примере ранних систем искусственного интеллекта (70ые годы 20 в): Student, Sir, Eliza. SHDRLU - Робот Винóграда – первая система КЛ с развитым лингвистическим компонентом. Метод шаблонов и продукционные правила. Понятие лингвистического процессора. Метод семантически ориентированного анализа А.С. Нариньяни как средство решения проблемы доступа на ЕЯ к структурированным источникам информации. Распознавание именованных сущностей.
4.	Машинный перевод (МП). Схемы МП. Анализ и синтез предложения. Понятие языка-посредника. Синтаксический анализ (отечественная традиция). Фильтровый метод.	Человеческий и машинный перевод (МП). Проблема несоответствия между языками. Типологические и контрастивные виды несоответствий между языками. Понятие языка-посредника. Прямая схема МП. Схема МП «интерлингва» и виды языков-посредников, понятия анализа и синтеза предложения. Схема МП «трансфер». Преимущества схемы «трансфер»

		перед схемой «интерлингва». Полнота синтаксического анализа и его применение в системах обработки текстов. Грамматика зависимостей как способ представления синтаксической структуры предложения. Фильтровые технологии.
5.	Формализмы синтаксического анализа в англоязычной традиции.	Метод шаблонов (Pattern matching), его применение в системах КЛ. Грамматика составляющих - Phrase Structure Grammar (PSG) - как способ представления синтаксической структуры предложения. Стратегии синтаксического анализа. Синтаксические формализмы: трансформационная грамматика Н. Хомского, Расширенные сети переходов (Augmented Transition Networks (ATN)); расширение и обобщение грамматики PSG: APSG и GPSG. Грамматики свойств и операция унификации. HPSG – Вершинная грамматика непосредственных составляющих.
Часть 2. Лингвистические ресурсы. Задачи и методы обработки языковых произведений – анализ и генерация текстов. Прикладные направления.		
6.	Корпуса текстов. Распространение эмпирических методов решения задач в КЛ. Метод n-gram – понятие статистической модели языка. Современное состояние МП.	Системы МП, основанные на примерах - example-based MT (EBMT), и системы типа «Переводческая память» (translation memory). Стадии обработки текста в системе МП, основанной на примерах (EBMT). Статистические системы МП и понятие статистической модели языка. Развитие и современное состояние прямых систем МП на примере системы ПРОМПТ. Развитие систем типа «трансфер» в сторону гибридных систем. Развитие систем, основанных на языке-посреднике, в сторону систем, основанных на знаниях (KBMT).
7.	Задачи семантики и типы семантических моделей в КЛ, связь с лексикографией. Лексико-семантические базы как технологии описания лексики.	Семантические типы и семантические примитивы. Падежные грамматики. Концептуализации в теории Р. Шенка, и семантическая классификация процессов. Классификация процессов по М. Хэллидею. Связь семантических типов процессов с семантическими ролями. Виды онтологий по особенностям описываемых понятий: философская, когнитивная, лексикографическая, лексическая, информационно-поисковые тезаурусы. Лексико-семантические отношения как средство описания парадигматического аспекта лексических систем: проект WordNet; принципы выделения отношений. Лексико-семантические базы на основе WordNet. Лексико-семантические базы как средство описания синтагматического аспекта лексических систем. Отечественные семантические словари. Проект FrameNet и иерархические отношения между фреймами.
8.	Автоматическая генерация текстов на ЕЯ. Теория риторических структур.	Виды входных данных для систем автоматической генерации текстов. Шаблонные и лингвистически мотивированные подходы. Схема

		идеализированной системы генерации текстов. Соответствие между типами входных данных и типами генерируемых текстов. Процессы автоматической генерации текстов: макро и микропланирование, лексикализация и грамматикализация, вставка ссылочных конструкций, агрегация, языковое оформление. Понятие дискурса и риторической структуры текста. Теория риторических структур (ТРС) как модель структуры дискурса текста: основные положения, определение риторического отношения, типы риторических отношений. Универсальный «план-оператор» для планирования текстов на основе определения риторического отношения.
9.	Общие модели структуры дискурса. Моделирование жанра и стиля текстов в ресурсах и системах КЛ.	Сравнение ТРС с другими моделями структуры дискурса: Теория связности дискурса Хоббса, модель структуры дискурса английского языка на основе системно-функциональной грамматики, моделирование методами формальной семантики – Discourse Representation Theory. Сравнение общих моделей дискурса. Понятие жанра. Классификация жанров в европейских корпусах текстов. Метаинформация о жанре текста в Национальном корпусе русского языка. Моделирование сюжета для текстов разных жанров: структура волшебных сказок (В.Я. Пропп); структура новеллы (В. Ленерт); структура специального текста (инструкции в системе генерации текстов AGILE). Понятия стиля в традиционной лингвистике. Попытки формализации понятия стиля в КЛ. Моделирование функциональных стилей в виде «конструкций» на основе статистической обработки текстов.
10.	Тематический аспект текста и поиск информации. Диалоговые и интерактивные системы.	Проявление тематического аспекта в текстах. Задачи информационного поиска: информационный поиск (ИП), фактографический поиск, поиск новой информации, распознавание именованных сущностей, отношений и тональности. Параметры полноты и точности ИП. Метод индексирования, основанный на частоте встречаемости слов как основа ИП. Метод кластеризации как способ классификации документов и терминов. Методы обработки неоднословных терминов. Использование информационно-поисковых тезаурусов. Лингвистические способы улучшения результатов статистических методов ИП.
11.	Автоматическая обработка устной речи и ее приложения.	Распознавание речи или преобразование «звук-буква»: преобразование речи в сеть измерений, преодоление вариаций, моделирование звуков речи, применение языковых ограничений, интеграция речевых и текстовых технологий. Синтез речи или преобразование «буква-звук». Статистические

		модели синтеза речи, основанные на звуковых корпусах. Применение речевых технологий.

4. Образовательные технологии

Для проведения учебных занятий по дисциплине используются различные образовательные технологии. Для организации учебного процесса может быть использовано электронное обучение и (или) дистанционные образовательные технологии.

5. Оценка планируемых результатов обучения

5.1 Система оценивания

Форма контроля	**Макс. количество баллов**	
	За одну работу	**Всего**
Текущий контроль:		
- домашние задания	*5 баллов*	*30 баллов*
- выполнение заданий на семинаре	*5 баллов*	*30 баллов*
Промежуточная аттестация – экзамен		40 баллов
Итого за семестр		100 баллов

Полученный совокупный результат конвертируется в традиционную шкалу оценок и в шкалу оценок Европейской системы переноса и накопления кредитов (European Credit Transfer System; далее – ECTS) в соответствии с таблицей:

100-балльная шкала	Традиционная шкала		Шкала ECTS
95 – 100	отлично	зачтено	A
83 – 94			B
68 – 82	хорошо		C
56 – 67	удовлетворительно		D
50 – 55			E
20 – 49	неудовлетворительно	не зачтено	FX
0 – 19			F

5.2 Критерии выставления оценки по дисциплине

Баллы/ Шкала ECTS	**Оценка по дисциплине**	**Критерии оценки результатов обучения по дисциплине**
100-83/ A,B	отлично/ зачтено	Выставляется обучающемуся, если он глубоко и прочно усвоил теоретический и практический материал, может продемонстрировать это на занятиях и в ходе промежуточной аттестации. Обучающийся исчерпывающе и логически стройно излагает учебный материал, умеет увязывать теорию с практикой, справляется с решением задач профессиональной направленности высокого уровня сложности, правильно обосновывает принятые решения. Свободно ориентируется в учебной и профессиональной литературе. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции, закреплённые за дисциплиной, сформированы на уровне –

Баллы/ Шкала ECTS	Оценка по дисциплине	Критерии оценки результатов обучения по дисциплине
		«высокий».
82-68/ C	хорошо/ зачтено	Выставляется обучающемуся, если он знает теоретический и практический материал, грамотно и по существу излагает его на занятиях и в ходе промежуточной аттестации, не допуская существенных неточностей. Обучающийся правильно применяет теоретические положения при решении практических задач профессиональной направленности разного уровня сложности, владеет необходимыми для этого навыками и приёмами. Достаточно хорошо ориентируется в учебной и профессиональной литературе. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции, закреплённые за дисциплиной, сформированы на уровне – «хороший».
67-50/ D,E	удовлетво-рительно/ зачтено	Выставляется обучающемуся, если он знает на базовом уровне теоретический и практический материал, допускает отдельные ошибки при его изложении на занятиях и в ходе промежуточной аттестации. Обучающийся испытывает определённые затруднения в применении теоретических положений при решении практических задач профессиональной направленности стандартного уровня сложности, владеет необходимыми для этого базовыми навыками и приёмами. Демонстрирует достаточный уровень знания учебной литературы по дисциплине. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции, закреплённые за дисциплиной, сформированы на уровне – «достаточный».
49-0/ F,FX	неудовлет-ворительно/ не зачтено	Выставляется обучающемуся, если он не знает на базовом уровне теоретический и практический материал, допускает грубые ошибки при его изложении на занятиях и в ходе промежуточной аттестации. Обучающийся испытывает серьёзные затруднения в применении теоретических положений при решении практических задач профессиональной направленности стандартного уровня сложности, не владеет необходимыми для этого навыками и приёмами. Демонстрирует фрагментарные знания учебной литературы по дисциплине. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции на уровне «достаточный», закреплённые за дисциплиной, не сформированы.

5.3 Оценочные средства (материалы) для текущего контроля успеваемости, промежуточной аттестации обучающихся по дисциплине

<u>В качестве домашних заданий предлагаются задачи следующих типов</u>

ДЗ1. Знакомство с организацией научно-практической области КЛ. Описание термина (по тезаурусу)
ДЗ2. Синтаксическая разметка фрагмента текста.
ДЗ3. Тестирование программы Элиза. Выделение видов шаблонов.
ДЗ4. Тестирование систем МП, организованных по разным схемам. Сравнение результатов.
(Проверка синтаксической разметки)
ДЗ5. Исследовательский проект (часть 1): корпусное исследование семантики русского глагола.
ДЗ6. Исследовательский проект (часть 2): Описание значений русского глагола в виде фреймов.

ДЗ7. Упражнение на построение модели дискурса конкретного текста в терминах теории риторических отношений.
(Проверка структуры текста. Обсуждение русских фреймов)
ДЗ8. Исследовательский проект (часть 3): Сопоставление русских фреймов с фреймами базы FrameNet.
ДЗ9. Тестирование современных диалоговых систем.
ДЗ10. Оформление отчета по исследовательскому проекту.

Экзамен ориентирован на следующие контрольные вопросы

КЛ как новая научно-практическая область исследований в контексте смежных наук.
Задачи теоретической КЛ и приложений.
Уровни языка и уровневые общие модели языка в КЛ.
Методы анализа текста.
Составляющие понятия «понимание». Понятие лингвистического процессора.
Семантически ориентированный метод анализа текстов.
Машинный перевод: схемы МП, понятие языка-посредника.
Синтаксический анализ предложения и фильтровый метод. Различия отечественной и англоязычной традиций в области синтаксического анализа.
Эволюция формализмов синтаксического анализа в англоязычной традиции.
Корпусы текстов и распространение эмпирических методов.
Типы систем МП, основанных на эмпирических методах. Понятие статистической «модели языка».
Падежные грамматики и семантические примитивы в КЛ. Связь синтаксиса и лексикологии.
Лексико-семантические базы как лексические ресурсы для систем обработки текстов.
Моделирование дискурса в ресурсах и системах КЛ.
Методы автоматической генерации текстов.
Моделирование параметров текстов (жанр, стиль) в ресурсах и системах КЛ.
Задачи поиска информации и мера терминологичности слова в тексте.
Виды поиска информации в КЛ.
Направления и приложения моделирования устной речи.

Планы семинарских занятий

Занятие 1(1)

Уровни языка в традиционной лингвистике и общие уровневые модели естественного языка (ЕЯ) в КЛ. Понятие лингвистического представления.

Цель занятия: На конкретных примерах показать различие принципов уровневого деления в традиционной лингвистике и общих моделях ЕЯ в КЛ. Ввести понятие лингвистического представления языкового объекта.

Занятие 2(1)

Задачи анализа предложения. Проблема неоднозначности.

Цель занятия: Познакомить студентов с системами КЛ, реализующими процессы «уровневого анализа»: сегментация, морфологический и синтаксический анализ, установление анафорических отношений. Сформировать у студентов представление о неоднозначности как специфической проблеме КЛ и познакомить с типами неоднозначности, возникающих при анализе на разных уровнях языка.

Занятие 3(2)

«Понимание» текстов в узкой предметной области. Метод шаблонов. Понятие лингвистического процессора.

Цель занятия: Познакомить студентов с моделированием понимания в узких предметных областях, выявляющим «информационные составляющие» процесса понимания. Продемонстрировать применение метода шаблонов. Дать понятие «лингвистического процессора».

Занятие 4(2)
Семантически ориентированный метод анализа. Распознавание именованных сущностей.

Цель занятия: Познакомить студентов с моделированием понимания в узких предметных областях и семантически-ориентированным методом А.С. Нариньяни как способами «обойти» проблему неоднозначности. Распознавание именованных сущностей как элемент семантически-ориентированного анализа текстов.

Занятие 5(3)
Машинный перевод (МП). Схемы МП. Анализ и синтез предложения. Понятие языка-посредника.

Цель занятия: Обсудить основы машинного перевода по сравнению с человеческим. Рассмотреть МП, основанный на правилах, как процесс преодоления несоответствий между языками и разрешения неоднозначности. Рассмотреть схему МП «интерлингва» и виды языков-посредников, UNL как амбициозный международный проект интерлингвы, понятия анализа и синтеза предложения. Преимущества схемы «трансфер» перед схемой «интерлингва».

Занятие 6(3)
Синтаксический анализ (отечественная традиция). Фильтровый метод.

Цель занятия: На примере синтаксического анализа познакомить студентов с фильтровым методом. Связать с проблемой неоднозначности в КЛ. Обсудить различие в эволюции синтаксического анализа в отечественных работах и за рубежом.

Занятие 7(4)
Формализмы синтаксического анализа в англоязычной традиции.

Цель занятия: Познакомить студентов с особенностями англоязычной традиции синтаксического анализа и кратко показать историю развития синтаксических формализмов от PSG до грамматик унификаций. Формализм как технология.

Занятие 8(5)
Корпуса текстов. Распространение эмпирических методов решения задач в КЛ. Метод n-gram – понятие статистической модели языка. Современное состояние МП.

Цель занятия: Системы МП, основанные на примерах - example-based MT (EBMT), и системы типа «Переводческая память» (translation memory). Стадии обработки текста в системе МП, основанной на примерах (EBMT). Статистические системы МП и понятие статистической модели языка. Развитие и современное состояние прямых систем МП на примере системы ПРОМПТ. Развитие систем типа «трансфер» в сторону гибридных систем. Развитие систем, основанных на языке-посреднике, в сторону систем, основанных на знаниях (KBMT).

Занятие 9(6)

Задачи семантики и типы семантических моделей в КЛ, связь с лексикографией.

Цель занятия: Семантические типы и семантические примитивы. Падежные грамматики. Концептуализации в теории Р. Шенка, и семантическая классификация процессов. Классификация процессов по М. Хэллидею. Связь семантических типов процессов с семантическими ролями.

Занятие 10(6)
Лексико-семантические базы как технологии описания лексики.

Цель занятия: Представить онтологии как средство организации понятий в КЛ и рассмотреть основные традиции описания понятий: философская, когнитивная, лексикографическая, лексическая, информационно-поисковые тезаурусы. Познакомить с основными технологиями описания парадигматических и синтагматических лингвистических отношений в проектах WordNet и FrameNet.

Занятие 11(7)
Автоматическая генерация текстов на ЕЯ.

Цель занятия: Обсудить виды входных данных для систем автоматической генерации текстов и подходы к моделированию выходных текстов, схему идеализированной системы генерации текстов, соответствие между типами входных данных и типами генерируемых текстов. Познакомить с процессами автоматической генерации текстов: макро и микропланирование, лексикализация и грамматикализация, вставка ссылочных конструкций, агрегация, языковое оформление.

Занятие 12(7)

Теория риторических структур.

Цель занятия: Понятие дискурса и риторической структуры текста. Теория риторических структур (ТРС) как модель структуры дискурса текста: основные положения, определение риторического отношения, типы риторических отношений.

Занятие 13(8)

Общие модели дискурса в КЛ.

Цель занятия: Обсудить общие модели структуры дискурса. Сравнение ТРС с другими моделями структуры дискурса: Теория связности дискурса Хоббса, модель структуры дискурса английского языка на основе системно-функциональной грамматики, моделирование методами формальной семантики – Discourse Representation Theory. Сравнение общих моделей дискурса.

Интернет-источники для аудиторной работы:
RST – (Rhetorical Structure Theory - Теория Риторических Структур): http://www.sfu.ca/rst/

Занятие 14(8)
Моделирование стиля и жанра текстов в ресурсах и системах КЛ.

Цель занятия: Обсудить моделирование понятий жанра для классификации текстов в европейских корпусах текстов и в НКРЯ, моделирование сюжета текстов разных жанров: волшебных сказок (В.Я. Пропп), новеллы (В. Ленерт), структуры специального текста - инструкции в системе генерации текстов AGILE. Представить лингвистическое понятие стиля и попытки формализации понятия стиля в КЛ, а также в системах генерации текстов. Попытки

моделирования функциональных стилей в виде «конструкций» на основе статистической обработки текстов.

Занятие 15(9)
Тематический аспект текста и поиск информации. Диалоговые и интерактивные системы.

Цель занятия: Выделить тематический аспект в текстах. Информационный поиск документов на основе метаинформации и автоматическая тематическая атрибуция текстов. Ввести понятие меры терминологичности tf*idf. Задачи информационного поиска: информационный поиск (ИП), фактографический поиск, поиск новой информации, распознавание именованных сущностей, отношений и тональности. Параметры полноты и точности ИП. Метод кластеризации как способ классификации документов и терминов. Лингвистические способы улучшения результатов статистических методов ИП. Применение диалоговых и интерактивных систем.

Занятие 16(10)
Автоматическая обработка устной речи.

Цель занятия: Сообщить основные принципы автоматической обработки устной речи, показать важность этого направления КЛ в современном обществе на примере приложений в быту, криминалистике, коммерции, производстве и т.д.

6. Учебно-методическое и информационное обеспечение дисциплины

6.1 Список источников и литературы

Основная литература

1. Боярский К.К. Введение в компьютерную лингвистику. Учебное пособие. – СПб ИТМО: 2013 – 73 стр. http://elib.ict.nsc.ru/jspui/bitstream/ICT/1452/1/1470.pdf
2. Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «ДИАЛОГ» - Москва, 2018 http://www.dialog-21.ru/media/4529/_-dialog2018scopus.pdf
3. Соколова Е.Г. Синтаксическая разметка в терминах грамматики зависимостей и синтаксических функций [Электронный ресурс]: метод. пособие. М.: РГГУ, 2011. -33 с. – Библиогр.: с. 33(5 назв.). – Режим доступа: http://elib.lib.rsuh.ru/elib/000003603.pdf
4. Chacon Thiago Costa. Improved computational models of sound change shed light on the history of the Tukanoan languages = Вопросы языкового родства / Thiago Costa Chacon, Johann-Mattis List // Вестник РГГУ. Серия "Филология. Вопросы языкового родства". - 2015. - № 3 (13). - С. 177-203. - Библиогр.: с. 200-203. - ил.
5. Church K.W., Mercer R.L. Introduction to the Special Issue on Computational Linguistics Using Large Corpora. // CL, vol. 19, 1993 – 24 p. https://aclanthology.info/pdf/J/J93/J93-1001.pdf
6. Stephen Hansen Michael McMahon Andrea Prat Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach The Quarterly Journal of Economics, Volume 133, Issue 2, 1 May 2018, Pages 801–870, https://doi.org/10.1093/qje/qjx045

Рекомендованная литература

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. Пособие /Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. – М.: МИЭМ, 2011. -272 с.

2. Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л. Лингвистический процессор для сложных информационных систем/ М.: Наука, 1992. 256 с.
3. Баранов А.Н. Введение в прикладную лингвистику. М., 2003. (Разделы: Моделирование общения.С. 20-25, и Моделирование структуры сюжета. С. 25-30.)
4. Болдасов М.В., Соколова Е.Г. Генерация текстов на естественном языке – теории, методы, технологии // НТИ, Серия 2, №7, 2006, с.1-15.
5. Болдасов М.В., Соколова Е.Г. Генерация текстов на естественном языке - состояние вопроса и прикладные системы // НТИ, Серия 2, №10, 2005, с.12-22.
6. Буторов В.Д. Моделирование синтаксиса естественного языка // Прикладное языкознание. Учебник. (ред. А.С.Гердт). СПб., 1996. С. 142-160.
7. Виноград Т. Система, понимающая естественный язык. М: Мир, 1976. (Разделы: Образец диалога. С. 21-32; Синтаксис и значение. С. 32-41; Базовый подход к представлению значений. С. 44-47.)
8. Дерюгина О. Программы-собеседники. НТИ серия 1, N 6, стр. 31-35.
9. Жигалов В.А., Соколова Е.Г. InBASE: технология построения ЕЯ интерфейсов к базам данных // Труды Международного семинара Диалог’2001 по компьютерной лингвистике и ее приложениям Том 2, Аксаково, Июнь 2001 с. 123-135. Доступна с сайта: http://dialog-21.ru/digest/archive/2001/?year=2001&vol=22725&id=6900
10. Зализняк А.А. Русский грамматический словарь, Изд. 2-е. М.: Рус. словарь, 2003.
11. Зализняк Анна А. Многозначность и смежные понятия. //Анна А. Зализняк. Многозначность в языке и способы ее представления. М: Языки славянских культур, 2006. Глава 1, 1.1. с. 20-34.
12. Искусственный интеллект. Справочник: В 3 кн. Кн. 1-2. М.: Радио и связь, 1990.
13. Кибрик А.А. Анализ дискурса в когнитивной перспективе. [Электронный ресурс] 2003 // http://www.philol.msu.ru/~otipl/new/main/people/kibrik-aa/files/DA_cognitive_perspective@Diss_2003.pdf
14. Кибрик, А. А. Модус, жанр и другие параметры классификации дискурсов // Вопросы языкознания. - 2009. - N 2. - С. 2-21.
15. Кобозева И.М. Лингвистическая семантика. Эдиториал УРСС. Москва. 2000. Раздел: Семантические валентности лексемы как семантические отношения, обусловленные ее лексическим значением. С. 140-146.
16. Леонтьева Н.Н. Автоматическое понимание текстов. Системы. Модели. Ресурсы. М.: Academia, 2006.
17. Лингвистический энциклопедический словарь. М., 1990.
18. Ляшевская О.Н., Кузнецова Ю.Л. Русский фреймнет: к задаче создания корпусного словаря конструкций // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27-31 мая 2009 г.). Вып. 8 (15). М.: РГГУ, 2009. С. 306-312.
19. Мельчук И.А. Об одной лингвистической модели типа «Смысл-Текст». Уровни представления языковых высказываний // Изв. АН СССР, Серия литературы и языка, 1974. Т. 33, №5-6. http://feb-web.ru/feb/izvest/1974/05/745-436.htm
20. Непейвода Н.Н. Квазиискусственный язык // Диалог’2002, Т 1, Москва: Наука. 2002. С.314-318. http://www.dialog-21.ru/materials/archive.asp?id=7361&y=2002&vol=6077.
21. Ножов И. Морфогическая и синтаксическая обработка текста (модели и программы)", 2003 год (диссертация). Глава 2. Доступна с сайта http://aot.ru/technology.html.
22. Соколова Е.Г. Синтаксическая разметка в терминах грамматики зависимостей и синтаксических функций [Электронный ресурс]: метод. пособие. М.: РГГУ, 2011. -33 с. – Библиогр.: с. 33(5 назв.). – Режим доступа: http://elib.lib.rsuh.ru/elib/000003603.pdf
23. Тестелец Я.Г. Введение в общий синтаксис. М.: Изд. центр РГГУ, 2001. С. 213-215, 722-747.
24. Цибульский В.В., Ежов А.С., Поляков Г.А., Феклистов В.В. Анализ и классификация времени и сроков в российских нормативных и правовых актах. // Компьютерная лингвистика и

интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 20012» (в печати).
25. Шаров С.А. Средства компьютерного представления лингвистической информации //1996. http://www.ksu.ru/eng/science/ittc/vol000/002/
26. Шенк Р. Обработка концептуальной информации. М.: Энергия, 1980.
27. Chris Manning and Hinrich Schütze, Chapter 7. Word sense disambiguation. Chapter 8. Lexical acquisition. Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA: May 1999. Доступна с сайта http://reslib.com/book/Foundations_of_Statistical_Natural_Language_Processing
28. Hutchins J. Machine translation: general overview // The Oxford handbook of computational linguistics (R. Mitkov ed.) N.Y.: Oxford university press, 2003. P. 501-511.
29. Jurafsky, Daniel, and James H. Martin. 2009. Chapter 10-12 in: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2nd edition. Prentice-Hall. Доступна с сайта http://lib.mexmat.ru/books/10138
30. Jurafsky, Daniel, and James H. Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2nd edition. Prentice-Hall. 2009.
31. Salton, Gerard. Automatic text processing. The transformation, analysis, and retrieval of information by computer. Addison-Wesley Publishing Company, Inc., 1989.
32. Survey of the state of the Art in Human Language Technology (Ronald Cole, editor in chief) Cambridge University Press. 1997. См. также: (http://cslu.cse.ogi.edu/HLTsurvey/).
33. The Oxford handbook of computational linguistics (R. Mitkov ed.) N.Y.: Oxford university press, 2003.
34. Uszkoreit H. What is computational linguistics? http://www.coli.uni-saarland.de/~hansu/what_is_cl.html

6.2 Перечень ресурсов информационно-телекоммуникационной сети «Интернет»

Интернет-источники:

А. Ресурсы:

- **Русско-английский тезаурус по компьютерной лингвистике**. Подобласть знаний: http://uniserv.iis.nsk.su/thes/index.php?ent=74.
- **Национальный корпус русского языка** (НКРЯ): http://www.ruscorpora.ru
  Синтаксический подкорпус НКРЯ: http://www.ruscorpora.ru/search-syntax.html
  Параллельный подкорпус НКРЯ:
- **Малый Академический Словарь** (МАС): http://feb-web.ru/feb/mas/mas-abc/default.asp.
- Лексико-семантические базы:
  WordNet (онтология значений полнозначных слов английского языка и лексико-семантические парадигматические отношения : http://wordnet.princeton.edu/ .
  Corelex: (типы регулярной полисемии английских существительных**) :** http://www.cs.brandeis.edu/~paulb/CoreLex/overview.html.
  FrameNet (значения и лексико-семантические синтагматические отношения английских предикатных слов в виде фреймов. Онтологические отношения фреймов) : http://framenet.icsi.berkeley.edu/.

Б. Системы анализа предложения по уровням:

- морфологический, синтаксический, поверхностный семантический:
  Dialing : http://aot.ru/
- семантико-синтаксический:
  HPSG (Вершинная грамматика непосредственных составляющих): http://www2.lingsoft.fi/cgi-bin/engcg?snt=Baby%2C+I+love+you.&h=on.
  Он-лайн демо: http://erg.emmtee.net.

LFG (Лексическая Функциональная Грамматика):
http://decentius.aksis.uib.no/logon/xle.xml

В. Прикладные системы онлайн:

- Программы-собеседники:
  Eliza («компьютерный психотерапевт») Дж. Вейценбаума:
  http://www.manifestation.com/neurotoys/eliza.php3.
- Вопросно-ответные системы:
  START (в MIT) : http://start.csail.mit.edu/
- Машинный перевод:
  ПРОМПТ (коммерческая, прямой): http://translate.ru
  ЭТАП3 (экспериментальная, трансфер): http://proling.iitp.ru/
  Dialing Translator (экспериментальная) : http://aot.ru/cgi-bin/translate.cgi
  GOOGLE. Translate (коммерческая, статистический):
  http://translate.google.com/translate_t#
- Анализ и поиск текстовой информации**)**: http://demo.rco.ru/
  RCO : http://demo.rco.ru/.

Г. Модели и теории (на Интернет-сайтах):

- **RST** – (Rhetorical Structure Theory - Теория Риторических Структур):
  http://www.sfu.ca/rst/.
- **UNL** (Universal Networking Language): http://www.undl.org/.

6.3 Профессиональные базы данных и информационно-справочные системы

Доступ к профессиональным базам данных: https://liber.rsuh.ru/ru/bases

Информационные справочные системы:

1. Консультант Плюс
2. Гарант

7. Материально-техническое обеспечение дисциплины

Для обеспечения дисциплины используется материально-техническая база образовательного учреждения: учебные аудитории, оснащённые компьютером и проектором для демонстрации учебных материалов.

№п/п	Наименование ПО	Производитель	Способ распространения
1	Adobe Master Collection CS4	Adobe	лицензионное
2	Microsoft Office 2010	Microsoft	лицензионное
3	Windows 7 Pro	Microsoft	лицензионное
7	Microsoft Share Point 2010	Microsoft	лицензионное
12	Windows 10 Pro	Microsoft	лицензионное
13	Kaspersky Endpoint Security	Kaspersky	лицензионное
14	Microsoft Office 2016	Microsoft	лицензионное
15	Visual Studio 2019	Microsoft	лицензионное
16	Adobe Creative Cloud	Adobe	лицензионное
17	Zoom	Zoom	лицензионное

8. Обеспечение образовательного процесса для лиц с ограниченными возможностями здоровья и инвалидов

В ходе реализации дисциплины используются следующие дополнительные методы обучения, текущего контроля успеваемости и промежуточной аттестации обучающихся в зависимости от их индивидуальных особенностей:

- для слепых и слабовидящих: лекции оформляются в виде электронного документа, доступного с помощью компьютера со специализированным программным обеспечением; письменные задания выполняются на компьютере со специализированным программным обеспечением или могут быть заменены устным ответом; обеспечивается индивидуальное равномерное освещение не менее 300 люкс; для выполнения задания при необходимости предоставляется увеличивающее устройство; возможно также использование собственных увеличивающих устройств; письменные задания оформляются увеличенным шрифтом; экзамен и зачёт проводятся в устной форме или выполняются в письменной форме на компьютере.
- для глухих и слабослышащих: лекции оформляются в виде электронного документа, либо предоставляется звукоусиливающая аппаратура индивидуального пользования; письменные задания выполняются на компьютере в письменной форме; экзамен и зачёт проводятся в письменной форме на компьютере; возможно проведение в форме тестирования.
- для лиц с нарушениями опорно-двигательного аппарата: лекции оформляются в виде электронного документа, доступного с помощью компьютера со специализированным программным обеспечением; письменные задания выполняются на компьютере со специализированным программным обеспечением; экзамен и зачёт проводятся в устной форме или выполняются в письменной форме на компьютере.

При необходимости предусматривается увеличение времени для подготовки ответа.

Процедура проведения промежуточной аттестации для обучающихся устанавливается с учётом их индивидуальных психофизических особенностей. Промежуточная аттестация может проводиться в несколько этапов.

При проведении процедуры оценивания результатов обучения предусматривается использование технических средств, необходимых в связи с индивидуальными особенностями обучающихся. Эти средства могут быть предоставлены университетом, или могут использоваться собственные технические средства.

Проведение процедуры оценивания результатов обучения допускается с использованием дистанционных образовательных технологий.

Обеспечивается доступ к информационным и библиографическим ресурсам в сети Интернет для каждого обучающегося в формах, адаптированных к ограничениям их здоровья и восприятия информации:

- для слепых и слабовидящих: в печатной форме увеличенным шрифтом, в форме электронного документа, в форме аудиофайла.
- для глухих и слабослышащих: в печатной форме, в форме электронного документа.
- для обучающихся с нарушениями опорно-двигательного аппарата: в печатной форме, в форме электронного документа, в форме аудиофайла.

Учебные аудитории для всех видов контактной и самостоятельной работы, научная библиотека и иные помещения для обучения оснащены специальным оборудованием и учебными местами с техническими средствами обучения:

- для слепых и слабовидящих: устройством для сканирования и чтения с камерой SARA CE; дисплеем Брайля PAC Mate 20; принтером Брайля EmBraille ViewPlus;
- для глухих и слабослышащих: автоматизированным рабочим местом для людей с нарушением слуха и слабослышащих; акустический усилитель и колонки;
- для обучающихся с нарушениями опорно-двигательного аппарата: передвижными, регулируемыми эргономическими партами СИ-1; компьютерной техникой со специальным программным обеспечением.

9. Методические материалы

9.1 Планы семинарских/ практических/ лабораторных занятий

1.	Введение. Краткая история и предмет компьютерной лингвистики (КЛ). Общая структура и основные задачи КЛ. Теоретическая и прикладная КЛ.	ДЗ1. Знакомство с организацией научно-практической области КЛ. Описание термина (по тезаурусу)
2.	Уровни языка в традиционной лингвистике и общие уровневые модели естественного языка (ЕЯ) в КЛ. Задачи анализа и синтеза предложения. Проблема неоднозначности.	ДЗ2. Синтаксическая разметка фрагмента текста.
3.	«Понимание» текстов в узкой предметной области. Метод шаблонов. Семантически ориентированный метод анализа.	ДЗ3. Тестирование программы Элиза. Выделение видов шаблонов.
4.	Машинный перевод (МП). Схемы МП. Синтаксический анализ (отечественная традиция). Фильтровый метод.	ДЗ4. Тестирование систем МП, организованных по разным схемам. Сравнение результатов.
5.	Формализмы синтаксического анализа в англоязычной традиции.	(Проверка синтаксической разметки)
6.	Корпуса текстов. Распространение эмпирических методов решения задач в КЛ. Метод n-gram – понятие статистической модели языка. Современное состояние МП.	ДЗ5. Исследовательский проект (часть 1): корпусное исследование семантики русского глагола.
7.	Задачи семантики и типы семантических моделей в КЛ, связь с лексикографией. Лексико-семантические базы.	ДЗ6. Исследовательский проект (часть 2): Описание значений русского глагола в виде фреймов.
8.	Автоматическая генерация текстов на ЕЯ. Теория риторических структур.	ДЗ7. Упражнение на построение модели дискурса конкретного текста в терминах теории риторических отношений. Проверка структуры текста. Обсуждение русских фреймов
9.	Общие модели дискурса. Моделирование стиля и жанра текстов в ресурсах и системах КЛ.	ДЗ8. Исследовательский проект (часть 3): Сопоставление русских фреймов с фреймами базы FrameNet.
10.	Тематический аспект текста и поиск информации. Диалоговые и интерактивные системы.	ДЗ9. Тестирование современных диалоговых систем.
11.	Автоматическая обработка устной речи и ее приложения.	ДЗ10. Оформление отчета по исследовательскому проекту.

9.2 Иные материалы

Все необходимые для обучения материалы даются на лекциях и практических занятиях.