

**Федеральное государственное автономное образовательное
учреждение высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»**

УТВЕРЖДЕНО

**Директор физтех-школы
прикладной математики и
информатики
А.М. Райгородский**

	Рабочая программа дисциплины (модуля)
по дисциплине:	Основы компьютерной лингвистики
по направлению:	Информатика и вычислительная техника
профиль подготовки:	Прикладная математика и информатика Физтех-школа Прикладной Математики и Информатики кафедра проблем передачи информации и анализа данных
курс:	2
квалификация:	магистр

Семестр, формы промежуточной аттестации: 3 (осенний) - Дифференцированный зачет

Аудиторных часов: 30 всего, в том числе:

лекции: 30 час.

семинары: 0 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 60 час.

Всего часов: 90, всего зач. ед.: 2

Программу составил: Л.Л. Иомдин, канд. филол. наук, старший научный сотрудник

Программа обсуждена на заседании кафедры проблем передачи информации и анализа данных 01.02.2024

Аннотация

Данный курс рассчитан на повышение осведомленности специализирующихся в компьютерной лингвистике, которая к настоящему времени сложилась как самостоятельная прикладная лингвистическая дисциплина. Специалист по деловой прозе обязательно должен иметь дело с современными компьютерными программами обработки естественно-языковых текстов, от простых программ типа корректоров орфографии до сложных систем автоматической обработки текстов, например, систем машинного перевода и автоматического информационного поиска, экспертных систем и других прикладных программ, которые, с одной стороны, моделируют искусственный интеллект в аспекте работы с обширными базами данных, а, с другой, требуют обязательных знаний современных аспектов компьютерной лингвистики, поскольку все эти системы базируются на естественном языке и на алгоритмах обработки языковых сведений.

1. Цели и задачи

Цель дисциплины

- познакомить магистрантов с важнейшими областями междисциплинарных исследований на стыке лингвистики со смежными дисциплинами, в первую очередь с компьютерной наукой.

Задачи дисциплины

- изучение основных математических моделей и результатов, используемых в математических методах анализа многомерных данных;
- научить магистрантов пользоваться методами обратной связи, т.е. применять полученные при разработке автоматических систем результаты для извлечения новых знаний о естественном языке;
- дать представление о месте теоретической лингвистики в задачах, решаемых компьютерной лингвистикой;
- познакомить магистрантов с современными подходами к решению задач автоматической обработки текстов, в том числе с гибридными и статистическими подходами и приемами машинного обучения.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
УК-2 Способен управлять проектом на всех этапах его жизненного цикла	УК-2.1 Формулирует в рамках обозначенной проблемы, цель, задачи, актуальность, значимость (научную, практическую, методическую и иную в зависимости от типа проекта), ожидаемые результаты и возможные сферы их применения
	УК-2.2 Способен прогнозировать результат деятельности и планировать последовательность шагов для достижения данного результата. Формирует план-график реализации проекта в целом и план контроля его выполнения
	УК-2.3 Способен организовать и координировать работу участников проекта, обеспечивать работу команды необходимыми ресурсами
	УК-2.4 Представляет публично результаты проекта (или отдельных его этапов) в форме отчетов, статей, выступлений на научно-практических конференциях, семинарах и т.п.
ПК-2.1	Знает основы научно-исследовательской деятельности в области информационных технологий, владеет знанием основ философии и методологии науки; знанием методов научных исследований и навыками их проведения

ПК-2 Понимает и способен применить в научно-исследовательской и прикладной деятельности основные законы естествознания, современный математический аппарат и алгоритмы, современные информационно-коммуникационные технологии	ПК-2.2 Умеет применять полученные знания в области фундаментальных научных основ теории информации и решать стандартные задачи в собственной научно-исследовательской деятельности
	ПК-2.3 Имеет практический опыт научно-исследовательской деятельности в области информационно-коммуникационных технологий
	ПК-2.4 Владеет методами и алгоритмами решения задач цифровой обработки сигналов, использования сети Интернет, аннотирования, реферирования, библиографического поиска, опыт работы с научными источниками
ПК-1 Готов к включению в профессиональное сообщество; способен проводить под научным руководством локальные исследования на основе существующих методов в конкретной области профессиональной деятельности	ПК-1.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации; владеет навыками подготовки научных обзоров, публикаций, рефератов и библиографий по тематике проводимых исследований на русском и английском языке
	ПК-1.2 Умеет решать научные задачи с пониманием существующих подходов к верификации моделей программного обеспечения в связи с поставленной целью и в соответствии с выбранной методикой
	ПК-1.3 Имеет практический опыт выступлений и научной аргументации при анализе объекта научной и профессиональной деятельности

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- основные цели и задачи компьютерной лингвистики;
- основные методы и подходы к автоматической обработке текстов (статистические, в т.ч. машинное обучение, гибридные);
- основные классы приложений, развиваемых на базе компьютерной лингвистики (информационный поиск, глубокий анализ данных, автоматический и автоматизированный перевод текстов с одного языка на другой, автоматическое аннотирование и реферирование документов, анализ тональности текста, человеко-машинное общение на естественном языке);
- основные классы цифровых лингвистических ресурсов, создаваемых методами компьютерной лингвистики (компьютерные одноязычные и многоязычные словари, аннотированные корпуса текстов).

уметь:

- строить базовые правила систем автоматической обработки текстов;
- разбираться в правилах и алгоритмах автоматической обработки текстов;
- строить базовые морфологические и синтаксические структуры предложения (на примере русского и английского языков).

владеть:

- навыком освоения большого объема информации;
- навыками постановки научно-исследовательских задач и навыками самостоятельной работы.

4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

4.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

№	Тема (раздел) дисциплины	Трудоемкость по видам учебных занятий, включая самостоятельную работу, час.			
		Лекции	Семинары	Лаборат. работы	Самост.

		лекции	семинары	лаборат. работы	работа
1	Лингвистика как наука о языке. Грамматика и словарь естественного языка.	6			12
2	Анализ и синтез текста.	6			12
3	Языковая неоднозначность. Правилловые и статистические подходы к автоматической обработке текста.	6			12
4	Система машинного перевода.	6			12
5	Обзор задач прикладной лингвистики.	6			12
Итого часов		30			60
Подготовка к экзамену		0 час.			
Общая трудоёмкость		90 час., 2 зач.ед.			

4.2. Содержание дисциплины (модуля), структурированное по темам (разделам)

Семестр: 3 (Осенний)

1. Лингвистика как наука о языке. Грамматика и словарь естественного языка.

Лингвистика как наука о языке. Представление об уровнях представления языка – фонетика, морфология, синтаксис, семантика. Лингвистика и прагматика.

Лингвистическое моделирование. Действующие модели языка. Теория «Смысл – Текст» как фундамент для построения систем автоматической обработки текста.

Грамматика и словарь естественного языка. Представление об интегральном описании языка. Представление о лексических функциях.

Краткий обзор формальных грамматик. Порождающие грамматики. Грамматики составляющих и грамматики зависимостей. Гибридные грамматики.

2. Анализ и синтез текста.

Анализ и синтез текста. Морфологический и синтаксический анализ. Парсинг. Различные подходы к синтаксическому анализу: анализ «сверху вниз» и «снизу-вверх».

3. Языковая неоднозначность. Правилловые и статистические подходы к автоматической обработке текста.

Языковая неоднозначность как принципиальное свойство языка и методы ее разрешения при автоматической обработке текста. Интерактивное разрешение лексической и синтаксической неоднозначности.

Правилловые и статистические подходы к автоматической обработке текста.

4. Система машинного перевода.

Задача машинного перевода в кругу задач автоматической обработки текста на естественном языке. Система машинного перевода как механизм обратной связи и источник новых лингвистических знаний.

Типы систем машинного перевода. Автоматический и автоматизированный перевод. Память переводов. Интерлингва (на примере UNL-универсального сетевого языка).

Морфологический компонент системы автоматической обработки текстов. Морфологическая структура слова и предложения.

Алгоритм синтаксического анализа. Синтаксические отношения. Синтагмы. Синтаксическая структура предложения.

Словарь системы автоматической обработки текстов. Словарь системы машинного перевода. Структура словарной статьи. Синтаксические признаки. Семантические признаки (дескрипторы). Теория валентностей. Модель управления.

Аннотированные корпуса текстов и их роль в задачах автоматической обработки текстов.

Синонимическое перефразирование высказываний и его прикладное значение.

5. Обзор задач прикладной лингвистики.

Современные цифровые лингвистические ресурсы (Word Net, Frame Net, Treebanks).

5. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

Аудитория с проектором.

6. Перечень рекомендуемой литературы

Основная литература

1. Начала компьютерной лингвистики [Текст] : уч. пособие для вузов / Ю. И. Шемакин .— М. : Изд-во МГОУ : Росвузнаука, 1992 .— 116 с.

Дополнительная литература

1. Прикладная и компьютерная лингвистика [Текст], коллективная монография/под ред. И. С. Николаева, О. В. Митрениной, Т. М. Ландо, -М., ЛЕНАНД, 2017

7. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

Не используются

8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень необходимого программного обеспечения и информационных справочных систем (при необходимости)

Программное обеспечение и информационные технологии не требуются.

9. Методические указания для обучающихся по освоению дисциплины (модуля)

Студент, изучающий дисциплину, должен с одной стороны, овладеть общим понятийным аппаратом, а с другой стороны, должен научиться применять теоретические знания на практике. В результате изучения дисциплины студент должен знать основные определения, понятия, аксиомы, алгоритмы.

Успешное освоение курса требует напряжённой самостоятельной работы студента. В программе курса приведено минимально необходимое время для работы студента над темой. Самостоятельная работа включает в себя:

- чтение и конспектирование рекомендованной литературы,
- проработку учебного материала (по учебной и научной литературе), подготовку ответов на вопросы, предназначенных для самостоятельного изучения, доказательство отдельных утверждений, свойств;
- подготовку к дифференцированному зачёту.

Руководство и контроль за самостоятельной работой студента осуществляется в форме индивидуальных консультаций.

Важно добиться понимания изучаемого материала, а не механического его запоминания. При затруднении изучения отдельных тем, вопросов, следует обращаться за консультациями к лектору.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению: Информатика и вычислительная техника
профиль подготовки: Прикладная математика и информатика
Физтех-школа Прикладной Математики и Информатики
кафедра проблем передачи информации и анализа данных
курс: 2
квалификация: магистр

Семестр, формы промежуточной аттестации: 3 (осенний) - Дифференцированный зачет

Разработчик: Л.Л. Иомдин, канд. филол. наук, старший научный сотрудник

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
УК-2 Способен управлять проектом на всех этапах его жизненного цикла	УК-2.1 Формулирует в рамках обозначенной проблемы, цель, задачи, актуальность, значимость (научную, практическую, методическую и иную в зависимости от типа проекта), ожидаемые результаты и возможные сферы их применения
	УК-2.2 Способен прогнозировать результат деятельности и планировать последовательность шагов для достижения данного результата. Формирует план-график реализации проекта в целом и план контроля его выполнения
	УК-2.3 Способен организовать и координировать работу участников проекта, обеспечивать работу команды необходимыми ресурсами
	УК-2.4 Представляет публично результаты проекта (или отдельных его этапов) в форме отчетов, статей, выступлений на научно-практических конференциях, семинарах и т.п.
ПК-2 Понимает и способен применить в научно-исследовательской и прикладной деятельности основные законы естествознания, современный математический аппарат и алгоритмы, современные информационно-коммуникационные технологии	ПК-2.1 Знает основы научно-исследовательской деятельности в области информационных технологий, владеет знанием основ философии и методологии науки; знанием методов научных исследований и навыками их проведения
	ПК-2.2 Умеет применять полученные знания в области фундаментальных научных основ теории информации и решать стандартные задачи в собственной научно-исследовательской деятельности
	ПК-2.3 Имеет практический опыт научно-исследовательской деятельности в области информационно-коммуникационных технологий
	ПК-2.4 Владеет методами и алгоритмами решения задач цифровой обработки сигналов, использования сети Интернет, аннотирования, реферирования, библиографического поиска, опыт работы с научными источниками
ПК-1 Готов к включению в профессиональное сообщество; способен проводить под научным руководством локальные исследования на основе существующих методов в конкретной области профессиональной деятельности	ПК-1.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации; владеет навыками подготовки научных обзоров, публикаций, рефератов и библиографий по тематике проводимых исследований на русском и английском языке
	ПК-1.2 Умеет решать научные задачи с пониманием существующих подходов к верификации моделей программного обеспечения в связи с поставленной целью и в соответствии с выбранной методикой
	ПК-1.3 Имеет практический опыт выступлений и научной аргументации при анализе объекта научной и профессиональной деятельности

2. Показатели оценивания компетенций

В результате изучения дисциплины «Основы компьютерной лингвистики» обучающийся должен:

знать:

- основные цели и задачи компьютерной лингвистики;
- основные методы и подходы к автоматической обработке текстов (статистические, в т.ч. машинное обучение, гибридные);
- основные классы приложений, развиваемых на базе компьютерной лингвистики (информационный поиск, глубокий анализ данных, автоматический и автоматизированный перевод текстов с одного языка на другой, автоматическое аннотирование и реферирование документов, анализ тональности текста, человеко-машинное общение на естественном языке);
- основные классы цифровых лингвистических ресурсов, создаваемых методами компьютерной лингвистики (компьютерные одноязычные и многоязычные словари, аннотированные корпуса текстов).

уметь:

- строить базовые правила систем автоматической обработки текстов;
- разбираться в правилах и алгоритмах автоматической обработки текстов;
- строить базовые морфологические и синтаксические структуры предложения (на примере русского и английского языков).

владеть:

- навыком освоения большого объема информации;
- навыками постановки научно-исследовательских задач и навыками самостоятельной работы.

3. Перечень типовых (примерных) вопросов, заданий, тем для подготовки к текущему контролю

С целью контроля освоения обучающимися учебного материала проводится устный опрос в начале или в конце занятия по пройденной теме.

4. Перечень типовых (примерных) вопросов и тем для проведения промежуточной аттестации обучающихся

Вопросы к дифференцированному зачёту:

1. Что такое уровни представления языковых выражений? Какие бывают уровни?
2. Морфологический анализ и синтез текстов. Поверхностная и глубинная морфология. Анализ композитов.
3. Основные типы представления синтаксической структуры предложения. Зависимости и составляющие. Дерево зависимостей.
4. Понятие синтаксического правила (синтагмы).
5. Грамматика и словарь.
6. Синтаксические признаки слова.
7. Валентностная структура предиката. Синтаксические и семантические валентности. Модель управления слова.
8. Основные типы компьютерных синтаксических ресурсов. Словари и корпуса текстов.
9. Глубокий анализ лингвистических данных: постановка задачи, основные методы и подходы.
10. Интерактивное разрешение лексической и синтаксической неоднозначности.

Критерии оценивания

Оценка отлично 10 баллов - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины, проявляющему интерес к данной предметной области, продемонстрировавшему умение уверенно и творчески применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений.

Оценка отлично 9 баллов - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений.

Оценка отлично 8 баллов - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, правильное обоснование принятых решений, с некоторыми недочетами.

Оценка хорошо 7 баллов - выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но недостаточно грамотно обосновывает полученные результаты.

Оценка хорошо 6 баллов - выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности.

Оценка хорошо 5 баллов - выставляется студенту, если он в основном знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач достаточно большое количество неточностей.

Оценка удовлетворительно 4 балла - выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он освоил основные разделы учебной программы, необходимые для дальнейшего обучения, и может применять полученные знания по образцу в стандартной ситуации.

Оценка удовлетворительно 3 балла - выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, допускающему ошибки в формулировках базовых понятий, нарушения логической последовательности в изложении программного материала, слабо владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и с трудом применяет полученные знания даже в стандартной ситуации.

Оценка неудовлетворительно 2 балла - выставляется студенту, который не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных принципов и не умеет использовать полученные знания при решении типовых задач.

Оценка неудовлетворительно 1 балл - выставляется студенту, который не знает основного содержания учебной программы дисциплины, допускает грубейшие ошибки в формулировках базовых понятий дисциплины и вообще не имеет навыков решения типовых практических задач.

5. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

Во время проведения дифференцированного зачёта обучающиеся могут пользоваться программой дисциплины, а также справочной литературой, вычислительной техникой.

Дифференцированный зачёт может проводиться по итогам текущей успеваемости и сдачи заданий, или путем организации специального опроса, проводимого в устной форме.