**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)**

| | |
|---|---|
| **Дисциплина:** | Основы компьютерной лингвистики |
| **Направление:** | Прикладные математика и физика |
| **Магистерская программа:** | Интеллектуальный анализ данных |
| | Физтех-школа прикладной математики и информатики |
| | Кафедра проблем передачи информации и анализа данных |
| **Курс:** | 2 |
| **Квалификация:** | Магистр |

Семестр, формы промежуточной аттестации: 3 (Осенний) – Экзамен

Аудиторных часов: 30 всего, в том числе:
лекции: 24 час.
практические (семинарские) занятия: 6 час.
лабораторные занятия: 0 час.

Самостоятельная работа: 15 час. всего, в том числе:
задания, курсовые работы: 0 час.

Подготовка к экзамену: 30 час.

Всего часов: 75, всего зач.ед.: 2

**Программу составил:** **Л.Л. Иомдин, кандидат филологических наук, доцент**

**Аннотация**

Лингвистика и компьютерная лингвистика. Лингвистика как наука о языке. Лингвистическое моделирование как задача компьютерной лингвистики. Действующие модели языка. Теория «Смысл – Текст» как фундамент для построения одной из таких моделей. Компьютерная лингвистика и задача автоматической обработки текста. Уровни представления языка – фонетика, морфология, синтаксис, семантика. Лингвистика и прагматика. Краткий обзор формальных грамматик как предтечи современной компьютерной лингвистики. Порождающая грамматика. Грамматики составляющих и грамматики зависимостей. Гибридные грамматики. Грамматика и словарь естественного языка. Представление об интегральном описании языка. Толково-комбинаторный словарь. Представление о лексических функциях.Автоматический анализ и синтез текста. Морфологический и синтаксический

анализ. Парсинг. Различные подходы к синтаксическому анализу: анализ «сверху вниз» и «снизу вверх». Языковая неоднозначность как принципиальное свойство языка и методы ее разрешения при автоматической обработке текста. Интерактивное разрешение лексической и синтаксической неоднозначности. Основные задачи автоматической обработки текста. Машинный перевод, извлечение информации, саммаризация, классификация текстов, логический вывод, определение тональности текстов. Правиловые и статистические подходы к автоматической обработке текста. 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). Правиловый, статистический и гибридный перевод. Современные системы машинного перевода на основе нейронных сетей. 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). Онтологические концепты. Теория валентностей. Модель управления. Аннотированные корпусы текстов и их роль в задачах автоматической обработки текстов. Синонимическое перифразирование высказываний и его прикладное значение. Современные цифровые лингвистические ресурсы (Word Net, Frame Net, Treebanks, аннотированные корпусы текстов). Методы дистрибутивной семантики. Векторно-пространственные модели определения семантической близости слов. Современные методы глубокого семантического анализа текста с участием лингвистических онтологий. Умозаключения на основе здравого смысла (common sense reasoning). Логический вывод.

**Содержание дисциплины (модуля), структурированное по темам (разделам)**

Семестр: 3 (Осенний)

1. Лингвистическое моделирование.

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.

2. Основные задачи и проблемы анализа естественно-языковых текстов.

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.
Анализ и синтез текста. Морфологический и синтаксический анализ. Парсинг. Различные подходы к синтаксическому анализу: анализ «сверху вниз» и «снизу вверх».
Языковая неоднозначность как принципиальное свойство языка и методы ее разрешения при автоматической обработке текста. Интерактивное разрешение лексической и синтаксической неоднозначности.
Правиловые и статистические подходы к автоматической обработке текста.
Алгоритм синтаксического анализа. Синтаксические отношения. Синтагмы. Синтаксическая структура предложения.

3. Машинный перевод и другие прикладные задачи компьютерной лингвистики.

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). Правиловый, статистический и гибридный перевод.
Морфологический компонент системы автоматической обработки текстов. Морфологическая структура слова и предложения.
Словарь системы автоматической обработки текстов. Словарь системы машинного перевода.
Структура словарной статьи. Синтаксические признаки. Семантические признаки (дескрипторы). Теория валентностей. Модель управления.
Аннотированные корпусы текстов и их роль в задачах автоматической обработки текстов.
Синонимическое перифразирование высказываний и его прикладное значение.
Обзор задач прикладной лингвистики.
Современные цифровые лингвистические ресурсы (Word Net, Frame Net, Treebanks).

4. Современные методы и средства глубокого семантического анализа текста.

Современные методы глубокого семантического анализа текста с участием лингвистических онтологий. Умозаключения на основе здравого смысла (common sense reasoning).

**Перечень основной и дополнительной литературы, необходимой для освоения дисциплины (модуля)**

Основная литература

1. Apresjan Ju, Boguslavsky I., Iomdin L et al. ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT // MTT 2003. First International Conference on Meaning – Text Theory (June 16-18, 2003). Paris: École Normale Supérieure, 2003. P. 279-288.
2. Мельчук И.А. Опыт теории лингвистических моделей «Смысл – Текст». М.: Языки славянской культуры, 1999. - 370 с.
3. Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л. и др. Лингвистический процессор для сложных информационных систем. М.: Наука, 1992. - 256 с.

Дополнительная литература

1. Boguslavsky I, Iomdin L. Nivre J. Parsing the Russian Dependency Treebank // Proceedings of COLING-2008. Manchester, 2008. P. 641-648.
2. Баранов А.Н. Введение в прикладную лингвистику // Серия «Новый лингвистический учебник». М.: Эдиториал УРРС. 2001. Глава 2, раздел 1.3.1. Моделирование общения (с. 20-31); Глава 4, разделы 1.3.1. – 1.3.4. «Естественный» перевод: лингвистические проблемы (с. 143-163); 1.4. Машинный перевод (с. 168-178).
3. Pollard C., Sag I.A. Head-Driven Phrase Structure Grammar // Chicago: University of Chicago Press. 1994. - 454 p.
4. Philipp Koehn. Statistical Machine Translation. Cambridge University Press, 2010. - 446 p.
5. Соснина Е.П. Введение в прикладную лингвистику // Ульяновск: УлГТУ, 2012. - 110 с.
6. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика // Учебное пособие. Большакова Е.И., Клышинский Э.С., Ландэ Д.В. и др. М.: МИЭМ, 2011. -272 с.
7. Jurafsky D., Martin J.H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition // Prentice Hall, 2009. - 988 p.