**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:** | Глубокое обучение в прикладных задачах 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:** | 1 |
| **Квалификация:** | Магистр |

Семестры, формы промежуточной аттестации: 2 (Весенний) - Дифференцированный зачет

Аудиторных часов: 30 всего, в том числе:
лекции: 15 час.
семинары: 15 час.
лабораторные занятия: 0 час.

Самостоятельная работа: 15 час.

Всего часов: 45, всего зач. ед.: 1

**Программу составил:** **А.А. Мовсесян**

**Аннотация**

Компьютерная лингвистика, с одной стороны, решает задачи, возникающие в разделе общей лингвистики, а с другой стороны, использует для их решения методы математического и компьютерного моделирования и искусственного интеллекта. В частности, на сегодняшний день широко применяются методы глубокого обучения. В то же время, прикладные задачи компьютерной лингвистики имеют приложение далеко за пределами собственно лингвистики: от обработки естественного языка и распознавания образов до построения рекомендательных систем. Данный курс знакомит студентов с основными прикладными задачами компьютерной лингвистики и современными методами их решения на основе машинного обучения. Полученные в рамках курса знания об этих методах и основных используемых архитектурах нейронных сетей помогут

студентам в их исследованиях в широком спектре задач искусственного интеллекта.

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

Семестр 2 (весенний)

1. Введение.

Лингвистика, компьютерная лингвистика, обработка естественного языка, машинное обучение, глубокое обучение. Взаимосвязь дисциплин. Введение в дистрибутивную семантику.

2. Дистрибутивная семантика.

Векторные представления слов. Word2vec. Методы снижения размерности векторных представлений. Матрица совместной встречаемости. Лексическая омонимия.

3. Основы нейронных сетей.

Полносвязная нейронная сеть. Нелинейность. Матричные вычисления. Стохастический градиентный спуск. Дифференцирование сложной функции. Якобиан. Метод обратного распространения ошибки.

4. Морфологическая разметка.

Морфология. Морфологический анализ и морфологическая разметка. Морфологический стандарт. Морфологическая омонимия. Морфологическая разметка как задача многоклассовой классификации.

5. Синтаксическая разметка.

Синтаксис. Грамматика составляющих и грамматика зависимостей. Синтаксическая омонимия. Синтаксическая разметка и синтаксический анализ. Проективность. Методы синтаксического анализа.

6. Рекуррентные нейронные сети.

Языковая модель. N-граммы. Языковая модель на основе нейронных сетей. Рекуррентная нейронная сеть. Исчезающий и взрывающийся градиент. LSTM и GRU. Двунаправленные и многослойные рекуррентные нейронные сети. Регуляризация. Дропаут.

7. Машинный перевод.

Статистический машинный перевод. Нейронный машинный перевод. Модель «последовательность в последовательность». Кодер и декодер. Жадные алгоритмы декодирования. Оценка качества автоматического перевода: метрика BLEU.

8. Механизм внимания.

Механизм внимания в модели «последовательность в последовательность». Мультипликативное внимание, низкоранговое мультипликативное внимание, аддитивное внимание. «Самовнимание». Нелинейность. Архитектура Трансформер. Позиционное кодирование. Кодер и декодер. «Многоголовое» и перекрестное внимание. Маски. Остаточные связи. Нормализация слоя.

9. Контекстные языковые модели.

Методы токенизации текста. Предобученные векторные представления слов. Предобучение как способ инициализации параметров модели. Языковая модель как метод предобучения. Дообучение. Методы предобучения. BERT и GPT как примеры предобученных моделей.

10. Семантический анализ.

Понимание естественного языка. Семантический анализ. Вопросно-ответные системы. Предметная область узкого и широкого профиля. Модель «Извлечение — Чтение».

11. Разрешение кореферентности.

Анафора и кореферентность с точки зрения лингвистики. Катафора. Методы разрешения кореферентности. Алгоритм Хоббса. Схема Винóграда. Бинарный классификатор и нейросетевые подходы.

12. Конвертация лингвистических данных.

Обучение без учителя. Машинный перевод как задача конвертации. Дуальное обучение. Многозадачное обучение. Конвертация морфологических стандартов. Отличия задачи разметки от задачи конвертации.

13. Интерпретируемость нейронных сетей
Нейронная сеть как черный ящик. Оценка качества как способ анализа модели. HANS. Использование языковых моделей для анализа нейронных сетей. Методы интерпретации «ответов» нейронной сети. Зашумление входных данных. Анализ «интерпретируемых» компонентов модели. Метод «прощупывания». Использование «неинтерпретируемых» моделей в правиловых системах.

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

Основная литература

1. Николенко С., Кадурин А., Архангельская Е. Глубокое обучение. Погружение в мир нейронных сетей. — СПб.: Питер, 2018. — 480 с.

Дополнительная литература

1. Manning C., Schutze H. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA, 1999. — 680 с.

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

- http://www.deeplearningbook.org (электронная версия книги Goodfellow I., Bengio Y., Courville A. Deep learning)
- https://web.stanford.edu/~jurafsky/slp3/ (черновик книги Jurafsky D., Martin J. H. Speech and Language Processing)
- https://docs.python.org/3/ (документация языка Python)
- https://pytorch.org/tutorials/ (руководство к библиотеке PyTorch)