Анализ естественного языка

Автор курса: **Литвинов Владимир Геннадьевич, к.т.н.**

Общая информация

Организационные вопросы

Объем курса: **144 часа (4 з.е.).**
Итоговая аттестация: **зачёт.**
Типы занятий:

1. Лекционные занятия - 16 часа (8 лекций).
2. Лабораторные работы - 24 часа (4 задания).
3. Самостоятельная работа - 102 часов.
4. И т.д.

Цели курса

1. Сформировать теоретические знания и практические навыки в области решения задач компьютерной лингвистики (ОЕЯ - обработки естественных языков, NLP - natural language processing) с использованием современных инструментов.
2. Разобраться с основными алгоритмами машинного обучения (ML - machine learning) для обучения систем на основе данных.
3. Познакомится с приложениями ОЕЯ.
4. Освоить современные методы ОЕЯ.

В рамках курса используются

1. Язык программирования **Python 3**.
2. Интерактивная среда **Jupyter Notebook**.
3. Библиотеки:
    - **numpy** - обработка массивов;
    - **pandas** - обработка признаковых описаний;
    - **matplotlib** - визуализация;
    - **scikit-learn** - алгоритмы ML;
    - **nltk** - алгоритмы ОЕЯ.
    - и другие.

Введение

История

1940-**е** - 50-**е**

- Изучением предмета знимаются множество разных наук.
- Первый машинный перевод (1954).
- Теория формальных языков.
- Теория информации.

История

1950-**е** - 70-**е**

- Публикация работы Хомского.
- ОЕЯ в рамках ИИ.
- Байесовские методы.
- Корпусная лингвистика.

История

1970-**е** - 80-**е**

- Статистические модели.
- Логическое представление языков.
- Переход к пониманию языка.
- Попытки перехода к дискурсу.

История

1980-**е** - 90-**е**

- Генерация речи.
- Конечные автоматы в морфологии.
- Решение data-driven задач (управления на основе данных).
- Новые способы оценки (evaluation).
- Вероятностные методы.

Что сейчас?

- Большое количество данных.
- Подходы “без учителя”.
- Обилие вычислительных ресурсов.
- Глубокое обучение.

Jeopardy game

(февраль 2011) - IBM Watson выиграла человека.

Голосовые помощники: начало

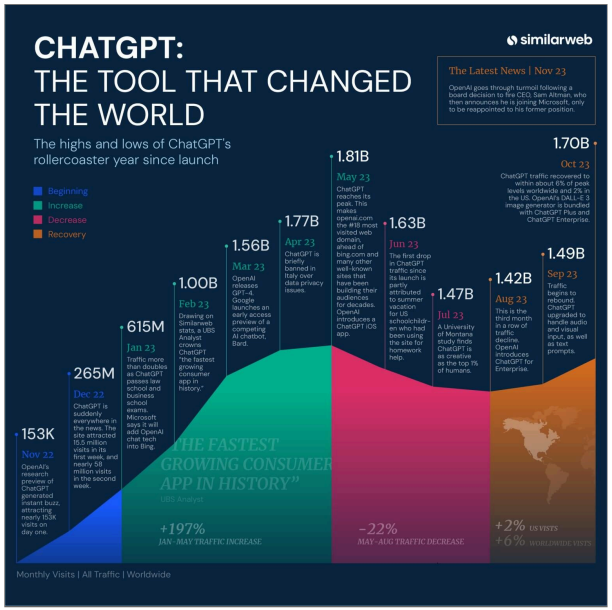
(октябрь 2011) - Siri.
(май 2016) - Google Assistant.
(1962) - IBM представила решение Shoebox.

Google Translate

(ноябрь 2016) - переход на нейронные сети (ANN).

ChatGPT (Generative Pre-trained Transformer)

(ноябрь 2022) - по заверениям создателей ChatGPT 5 пройдет тест Тьюринга.



Цель ОЕЯ

- Основная цель - научить компьютер понимать естественные языки и общаться на них.
- В реальности:
    - лингвистические задачи;
    - information extraction: выделение информации из текста;
    - information retrieval: выдача информации по запросу;
    - работа со звуковым рядом: распознавание/синтез речи,
    - смежные задачи;
    - прочее...

Задачи ОЕЯ

- категоризация/классификация текстов;
- тематическое моделирование (topic modeling);
- машинный перевод (machine translation);
- ответ на вопросы (question answering, не путать с information retrieval);
- анализ эмоциональной окраски (sentiment analysis);
- автоматическое реферирование (summarization);
- выделение именованных сущностей (named entity recognition);
- разбор синтаксиса (как построить синтаксическое дерево предложения);
- разделение различных смыслов слов (word sense disambiguation);
- разбор морфологии, анафора, части речи, ...

Почему ОЕЯ - это сложно?

Вспомним классификацию по Хомскому формальных языков и грамматик

Грамматика языка - это множество объектов: $G = \langle V_T, V_N, R, S \rangle$.

- V_T - множество терминальных символов.
- V_N - множество нетерминальных символов.
- R - множество правил.
- S - аксиома.

A-грамматики

Грамматика G называется грамматикой типа 3, регулярной, праволинейной или автоматной грамматикой (А-грамматикой), если каждое правило из R имеет вид: $A \to xB$ (праволинейное правило) или $A \to x$ (заключительное правило), где $A, B \in V_N$, $x \in V_T$.

КС-грамматики

Грамматика G называется грамматикой типа 2, бесконтекстной или контекстно-свободной (КС-грамматикой), если ее правила имеют вид: $A \to \alpha$, где $A \in V_N$, $\alpha \in (V_N \cup V_T)^*$.

КЗ-грамматики

Грамматика G называется грамматикой типа 1, контекстной, нормальных составляющих (НС-грамматикой) или контекстно-зависимой (КЗ-грамматикой), если ее правила имеют вид: $\varphi A \psi \to \varphi \alpha \psi$, где $A \in V_N$, φ, $\psi \in (V_N \cup V_T)^*$ и $\alpha \in (V_N \cup V_T)^+$.

Грамматики с фразовой структурой

Грамматика G называется грамматикой типа 0, грамматикой с фразовой структурой или рекурсивно-перечислимой грамматикой, если ее правила имеют вид: $\alpha \to \beta$, где на левую и правую части правил не наложено никаких ограничений.

В чем еще заключены сложности?

- Естественные языки - это не языки программирования. Они не проектируются, они развиваются самостоятельно:
    - постоянно появляются новые слова;
    - крайне затруднителен синтаксический анализ;
    - присуща неоднозначность (ambiguity).
- Для интерпретации необходимы знания о окружающем мире.
- Большое количество языков, диалектов, стилей и т.д.

Терминология

- **Изолирующий язык** - язык характеризующийся неизменяемостью слов (отсутствие форм словоизменения) и выражение синтаксических отношений преимущественно посредством порядка слов. Примером может служить китайский язык.
- **Флективный язык** - язык использующий в морфологии главным образом флексию. Пример: "пол-е", "пол-я", "пол-ей".
- **Агглютинативный язык** — языки, имеющие строй, при котором доминирующим типом словоизменения является агглютинация («приклеивание») различных формантов (суффиксов или префиксов). Пример: die Küche + der Tisch = der Küchentisch.

Терминология

- **Полисемия** (многозначность) - это способность одного слова служить для обозначения разных предметов и явлений действительности (ассоциативно связанных между собой).
    - Лексическая полисемия. Пример: "Комбайн на поле.", "Электрическое поле." - одно слово служит для обозначения нескольких предметов/явлений.
    - Грамматическая полисемия. Пример: "Позвонили в колокол." (глагол в неопределенно-личном значении), "Мы позвонили Васе." (глагол в собственно-личном значении) - слово (обычно глагол) можно употребить в нескольких значениях.

Терминология

- **Омонимия** - совпадение по форме двух разных по смыслу единиц.
    - Лексическая омонимия. Пример: "Древний замок.", "Дверной замок." - слова, относящиеся к одной и той же части речи и имеющие одинаковые леммы, различаются лишь лексическим смыслом.
    - Морфологическая омонимия. Пример: "Данная история отложилась в веках.", "Снег медленно таял на её веках." - одно и то же слово является одной частью речи, но относится к разным леммам и совпадает лишь в некоторых формах.
    - Частеречная омонимия. Пример: "Моя пол.", "Моя машина."- одно и то же слово относится к различным частям речи.

Разделы ОЕЯ

- **Фонология** – изучает звуки речи и правила их соединения при формировании речи.
- **Лексикография** описывает лексикон конкретного ЕЯ (естественного языка), его отдельные слова и их грамматические свойства, а также методы создания словарей.
- ***Морфология*** – занимается внутренней структурой и внешней формой слов речи, включая части речи и их категории.
- ***Синтаксис*** – изучает структуру предложений, правила сочетаемости и порядка следования слов в предложении, а также общие его свойства как единицы языка.
- ***Семантика* и прагматика** – тесно связанные области: семантика занимается смыслом слов, предложений и других единиц речи, а прагматика – особенностями выражения этого смысла в связи с конкретными целями общения.

Уровни разбиения текста

- **Уровень предложений** (высказываний) – синтаксический уровень;
- **Уровень слов** (словоформ – слов в определенной грамматической форме) – морфологический уровень;
- **Уровень фонем** (отдельных звуков, с помощью которых формируются и различаются слова) – фонологический уровень