Лекция 1

Введение в обработку естественного языка

Логистика курса

- Инструктор: Антюхов Денис Олегович
- Занятия: сб и вс, Большой Левшинский 1/11, с 10 до 14
- Материалы: https://github.com/gaphex/mcs-nlp/
- Slack: mcs-chat.slack.com
- Д/З каждую неделю, принимаются с помощью telegram-ботов
- В конце курса - соревнование ботов

План лекции (теория)

- Обработка естественного языка: определение, цели
- Проблемы неоднозначности в NLP
- Исторический экскурс
- Состояние технологий на сегодняшний день
- Приложения NLP в индустрии

План лекции (практика)

- Создание нового телеграм-бота, получение токена
- Внутреннее устройство бота
- Запуск
- Первое домашнее задание

Определение

Обработка естественного языка - область знаний на стыке

- компьютерных наук
- лингвистики
- искусственного интеллекта

NLP занимается проблемами анализа и синтеза языка, на котором общаются люди

Цель

Цель: умная обработка естественных языков

- Разработка машин, способных понимать человеческие языки с тем чтобы взаимодействовать с людьми удобным образом
- Создание технологий для решения прикладных задач (информационный поиск, распознавание речи,)

В чем сложность?

“At last, a computer that understands you like your mother”
(Реклама ASR, 1990)

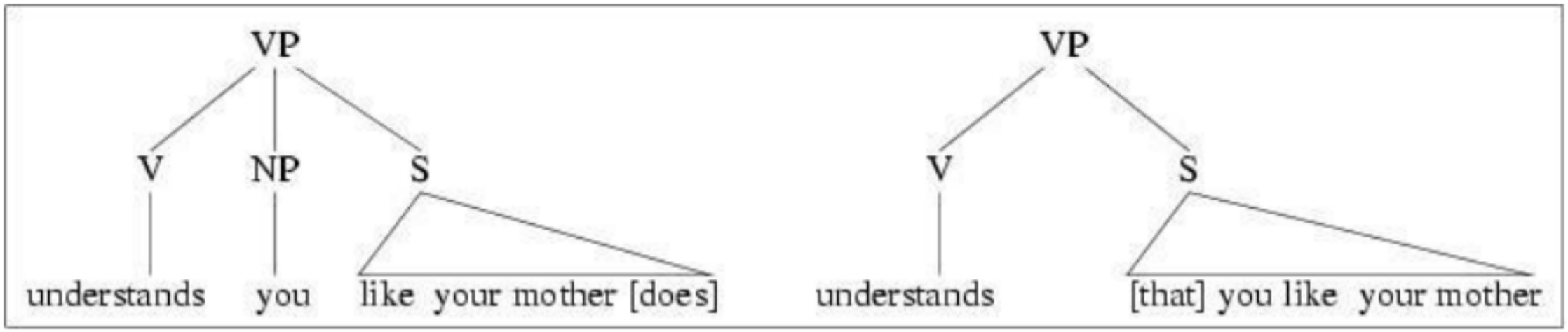
1. The computer understands you as much as your mother understands you.

2. The computer understands that you like your mother.

3. The computer understands you as well as it understands your mother.

В NLP остро стоит проблема неоднозначности

Синтаксическая неоднозначность



Семантическая неоднозначность

- mother - родитель женского пола

- mother - вязкое волокнистое вещество, состоящее из дрожжевых клеток и бактерий; добавляется в сидр или вино для производства уксуса (Oxford Dictionary)

Дискурсная неоднозначность

“Мы отдали бананы обезьянам, потому что они были голодные”

“Мы отдали бананы обезьянам, потому что они были перезрелые”

Правильное понимание смысла зависит от наших знаний об обезьянах и бананах

Свободный порядок слов

“Бытие определяет сознание”

Для решения задач NLP машине необходимы знания о

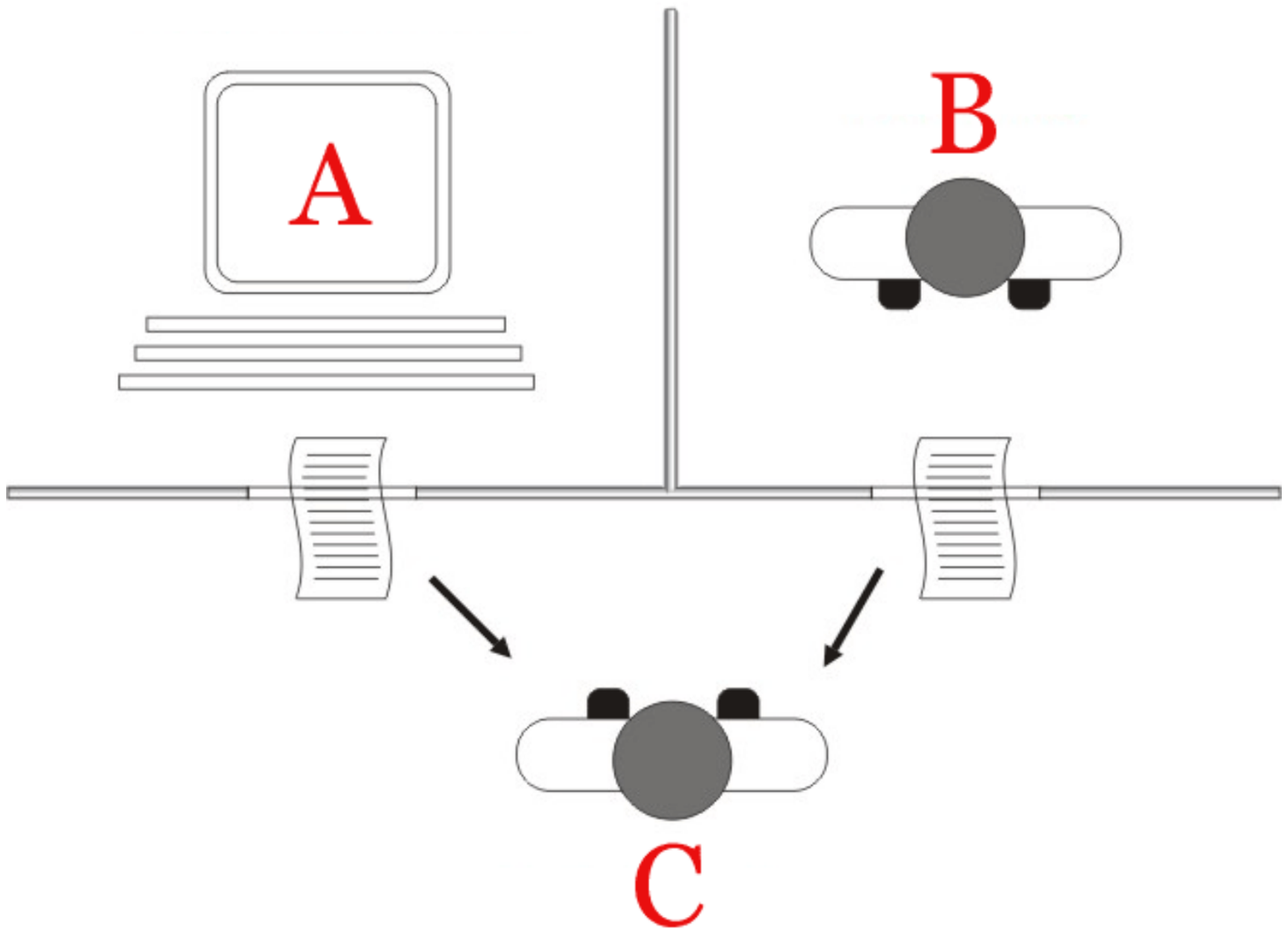
- правилах языка (лингвистике)
- окружающем мире

Исторически сложилось две парадигмы

- **Символический подход:** при помощи экспертов, вручную закодировать всю необходимую лингвистическую информацию в компьютер
- **Эмпирикеский подход:** имея достаточное количество данных, вывести свойства и правила языка из них

Краткая история NLP

- 1950 год
  Алан Тьюринг публикует статью “Computing Machinery and Intelligence” в которой предлагает т.н. Тест Тьюринга в качестве критерия наличия интеллекта



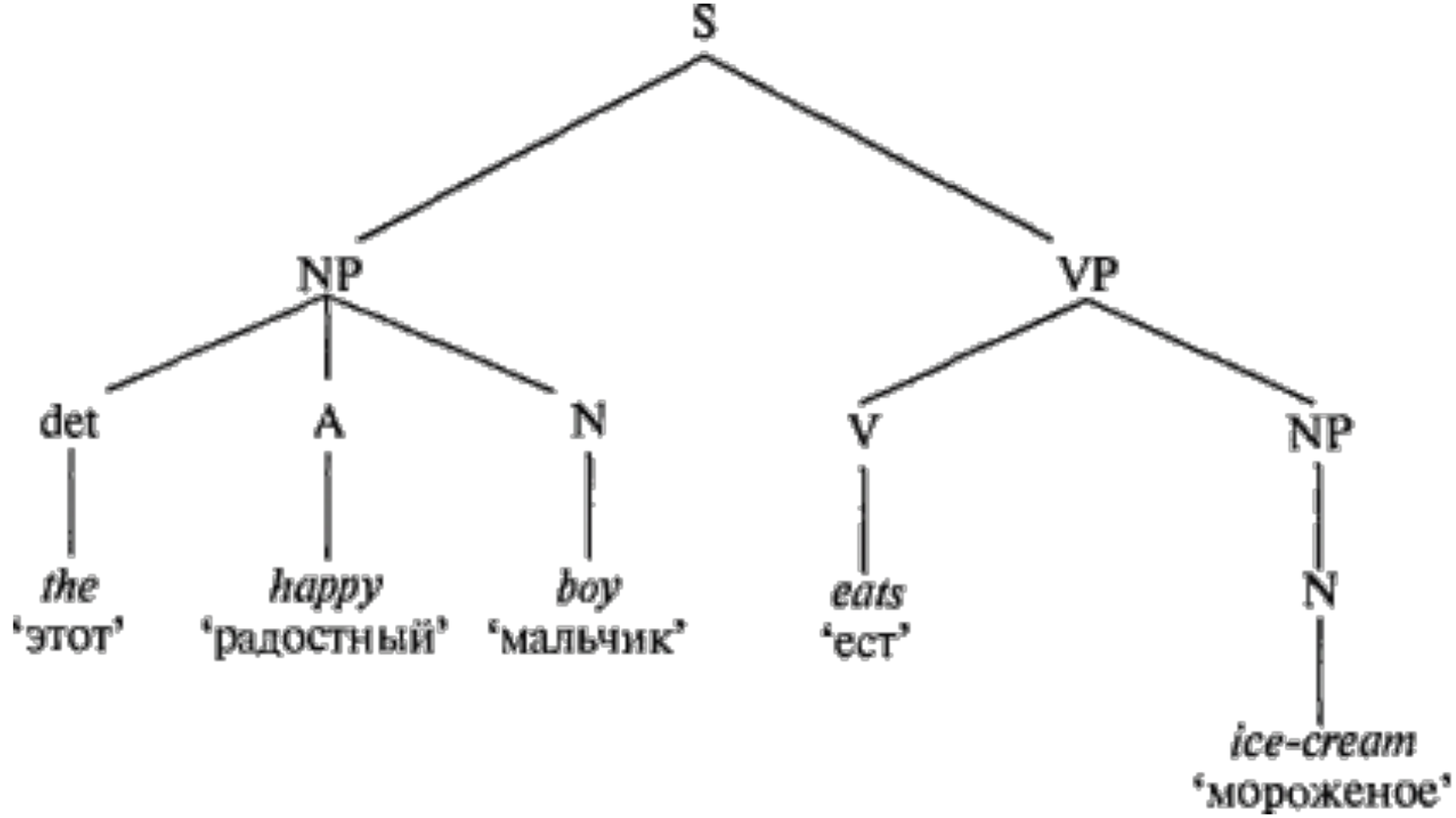
- 1954, Джорджтаун.
  С помощью мейнфрейма IBM-701 получен автоматический перевод 60 русских предложений на английский язык. Исследователи утверждают, что задача машинного перевода будет решена в течение следующих трех - четырех лет.

Colourless green ideas sleep furiously
Furiously sleep ideas green colourless

It is fair to assume that neither sentence (1) nor (2) has ever occurred in an English discourse. Hence, in any computed statistical model these sentences will be ruled out on identical grounds as equally "remote" from English. Yet (1), though nonsensical, is grammatical, while (2) is not.

Noam Chomsky, 1957

- 1956 - 1958, Нью-Хэмпшир.
  Ноам Хомски публикует книгу «Синтаксические структуры», в которой он предлагает т.н. грамматику составляющих адаптированную для обработки текста компьютером.

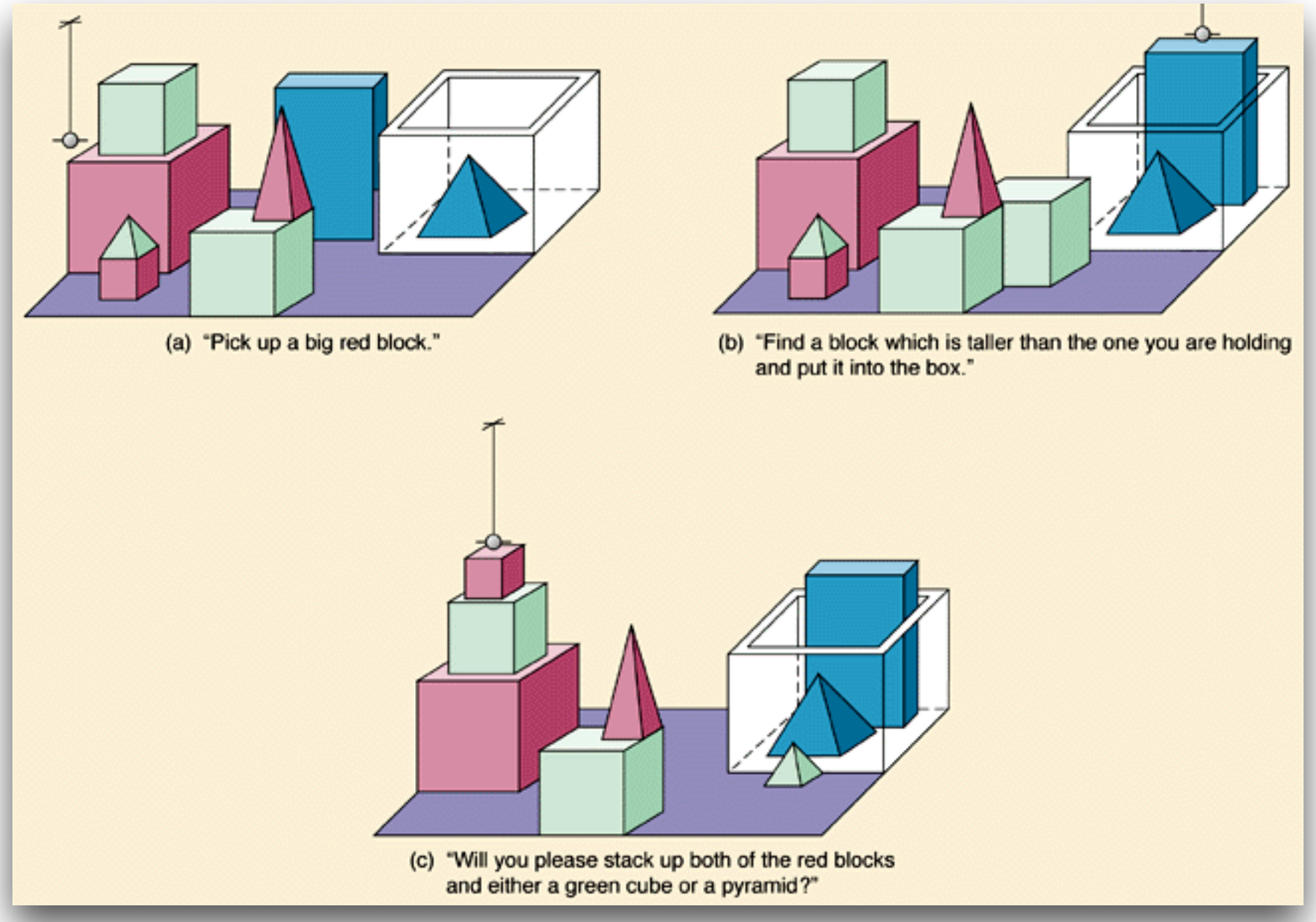


- 1964, США
  Для оценки хода исследований в области NLP основан Консультативный комитет по автоматической обработке естественно го языка (Automatic Language Processing Advisory Committee, ALPAC).

- 1966, США
  ALPAC публикует отчет, в котором приходит к выводу что «машинный пере вод общего научного текста реализован не был и не будет реализован в ближайшей перспективе».

John Robinson Pierce, Head of ALPAC

- 1970, Массачусетс
  Терри Виноградом представляет проект SHRDLU который позволяет пользователю взаимодействовать с виртуальным "блочным миром"содержащим различные объекты: блоки, конусы, шары, при помощи английских терминов. Проект имеет оглушительный успех как демонстрация возможностей искусственного интеллекта.



Внутренности SHRDLU

```
(DEFTHEOREM TC-GRASP
          (THCONSE (X Y)    (#GRASP $?X)
                    (THGOAL(#MANIP $?X))
                    (THCOND ((THGOAL (#GRASPING $?X)))
                             ((THGOAL (#GRASPING $_Y))
                              (THGOAL (#GET-RID-OF $?Y)
                                      (THUSE TC-GET-RID-OF))))
                             (T))
                    (THGOAL (#CLEARTOP $?X) (THUSE TC-CLEARTOP))
                    (THSETQ $_Y (TOPCENTER $?X))
                    (THGOAL (#MOVEHAND $?Y)
                             (THUSE TC-MOVEHAND))
                    (THASSERT (#GRASPING $?X))))

(DEFTHEOREM TC-PUT
          (THCONSE (X Y Z)  (#PUT $?X $?Y)
                    (CLEAR $?Y (SIZE $?X) $?X)
                    (SUPPORT $?Y (SIZE $?X) $?X)
                    (THGOAL (#GRASP $?X) (THUSE TC-GRASP))
                    (THSETQ $_Z (TCENT $?Y (SIZE $?X)))
                    (THGOAL (#MOVEHAND $?Z)  (THUSE TC-MOVEHAND))
                    (THGOAL (#UNGRASP)  (THUSE TC-UNGRASP))))
```

Требуется подробное описание лингвистических правил в коде

- 1974 год
  Предложен эффективный алгоритм обратного распространения ошибки (backpropagation)

- 1980 год
  Эмпирическая революция в NLP
  Возросшие вычислительные мощности позволили применять методы статистического моделирования и машинного обучения к задачам NLP

- 1984 год
  Крупные успехи группы IBM в автоматическом машинном переводе ознаменовывают переход к статистической парадигме в NLP

"Every time I fire a linguist, the performance of the speech recognizer goes up"
Fred Jelinek, 1985

Наши дни

с ростом вычислительных мощностей и доступных объемов данных, эмпирические подходы постепенно вытеснили символьные из всех областей NLP

- 2013 - word2vec, плотные семантические вектора слов
- 2014 - IBM Watson обрабатывает клиническую историю пациентов для определения диагнозов
- 2015 - Диалоговые агенты основанные на глубоком обучении (Siri, Alexa) доступны на мобильных устройствах
- 2016 - Google Neural Machine Translation работает с >100 языковых пар
- 2017 - Яндекс внедряет глубокое обучение для ранжирования поисковой выдачи («Королев»)

NLP в индустрии

- Информационный поиск
  поиск по ключевым словам, синонимам

- Вопросное-ответные системы
  чат-боты, агенты поддержки клиентов

- Классификация текста
  сентиментный анализ

- Машинный перевод

- Чат-боты
  диалоговые-агенты, вопросное-ответные системы,
  службы поддержки клиентов

Состояние NLP на сегодняшний день

mostly solved

Spam detection

Let's go to Agra! ✓

Buy V1AGRA ... ✗

Part-of-speech (POS) tagging

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.

Named entity recognition (NER)

PERSON ORG LOC

Einstein met with UN officials in Princeton

making good progress

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕...

The 13th Shanghai International Film Festival...

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

⇨ Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

Практическая часть

Клонируем репозиторий

- Ставим git
  https://git-scm.com/book/en/v2/Getting-Started-Installing-Git

- В консоли выполняем
  git clone https://github.com/gaphex/mcs-nlp/

- Переходим в директорию lecture1
  cd lecture1

Регистрируем нового бота

- устанавливаем Telegram если еще не
- переходим на https://telegram.me/BotFather
- либо просто пишем боту @BotFather
- команда /help даст справку
- вводим команду /newbot
- придумываем имя и юзернейм
- получаем токен для нового бота

Конфигурируем и запускаем

- редактируем файл lecture1/config.py

- пишем свой токен в переменную TOKEN

- убеждаемся что установлен пакет python-telegram-bot
  если нет, то
  pip3 install python-telegram-bot - - user или
  conda install python-telegram-bot

- запускаем бота!
  python3 bot.py

Домашнее задание

- используя модели полученные в ходе соревнования на kaggle, реализовать добавить в своего бота функциональность сентиментного анализа

- вызов модели сделать внутри функции get_sentiment

- бот должен обрабатывать команды вида /sentiment This is a sample message