Intro to NLP

Lecture Plan

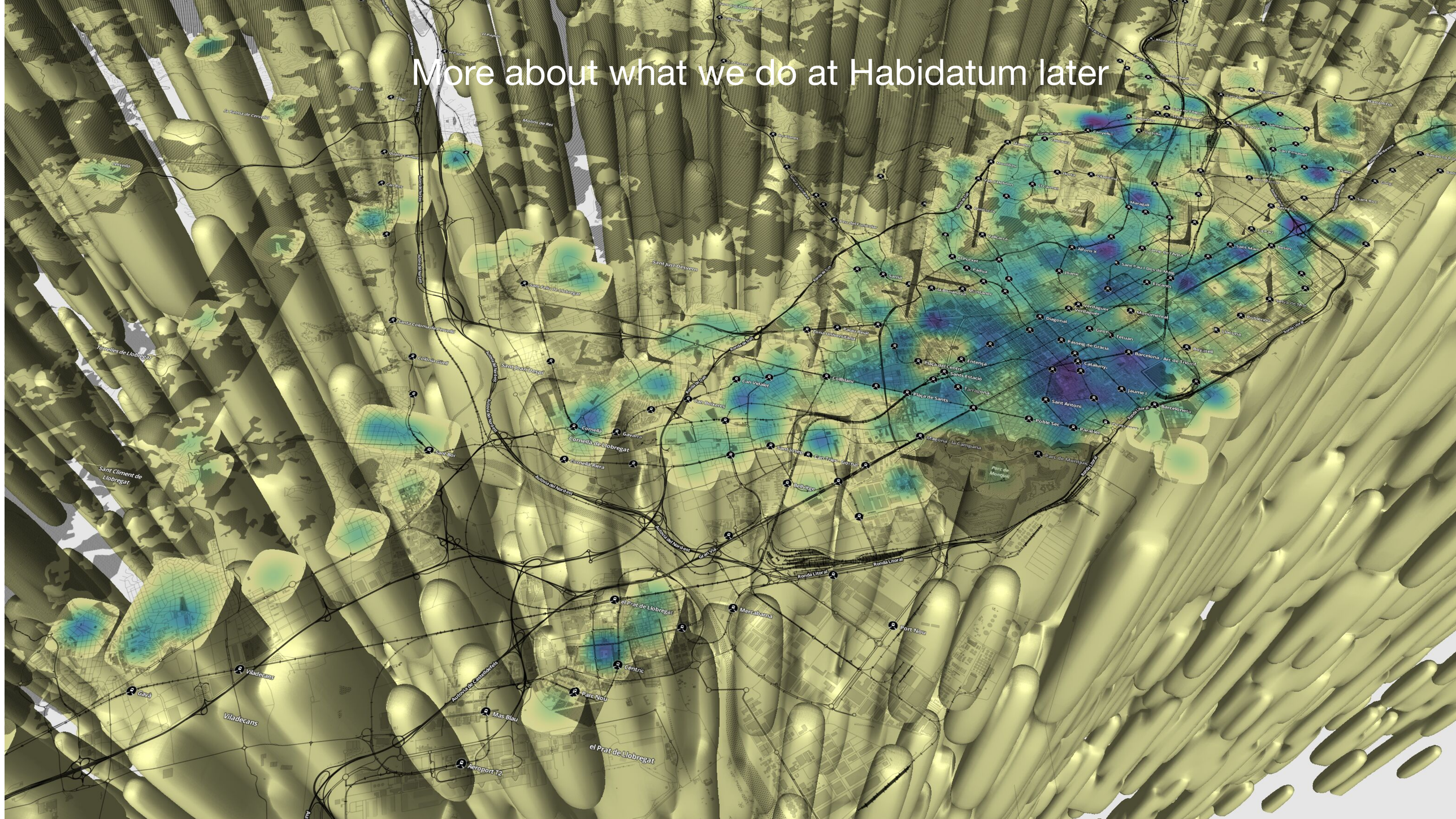
- NLP: goals and definitions
- Ambiguity problems in NLP
- Some history
- NLP today
- Some NLP applications

Practical Plan

- Let’s create a telegram bot
- Bot registration, set up
- How the bot works
- Using a trained model for Sentiments
- Basic text processing with spaCy

About the lecturer

- Nikita Pestrov, ipestrov@gmail.com
- Data Science Lead at Habidatum
- We analyse urban data and focus on space-time patterns
- MIPT BA 2014, Skoltech MS 2016, Thesis on pattern mining at MIT Media Lab
- easy ten co-founder


More about what we do at Habidatum later

Definitions

Natural Language Processing is an are on a merge of

- linguistics
- computer science
- artificial intelligence

NLP is focused on human language analysis and language synthesis.

NLP Goal

- Intellectual analysis of natural languages
- Development of machines that can understand human language and interact with them more naturally
- Development of technologies to solve practical tasks (translation, entity extraction, etc.)

What is so hard about it?

“At last, a computer that understands you like your mother”
(1985 McDonnell-Douglas ad)

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

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

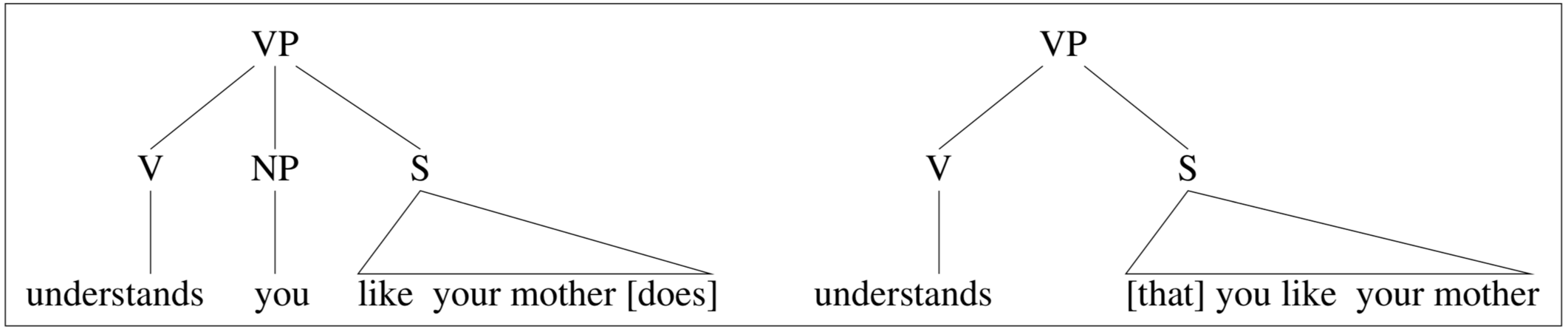
3. The computer understands that you like your mother.

The issue of ambiguity is very noticeable at NLP.

Audio level

- A computer that understand you like your mother
- ..A computer that understands your lie cured mother

Syntactical level



- Different structure => Different interpretation

Semantic (meaning) level

Two definitions of “mother”

- a woman who has given birth to a child
- a stringy slimy substance consisting of yeast cells and bacteria; is added to cider or wine to produce vinegar

Word sense ambiguity

How to solve this?

NLP solutions need:

- Knowledge about language
- Knowledge about the world

Possible approaches:

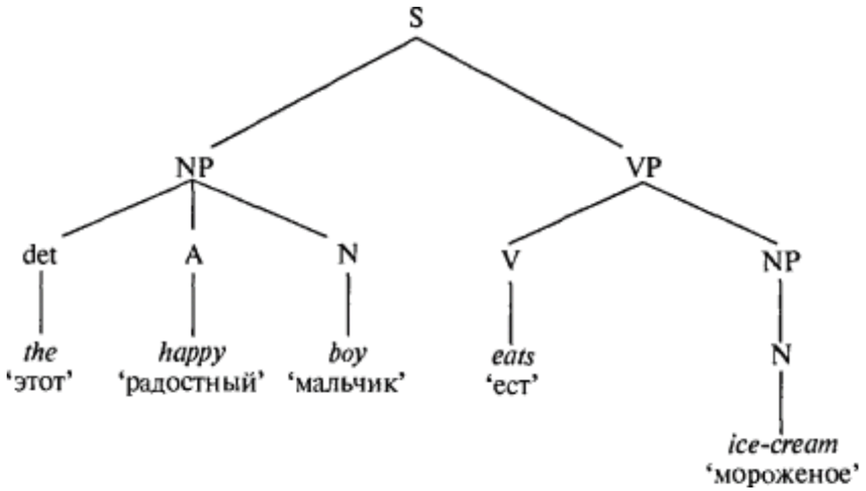
- Symbolic approach: Encode all the required information into computer
- Statistical approach: Infer language properties from language samples

The era of symbolic NLP

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

In 1956-1958 Noah Chomsky publishes a book "Syntactic structures", with a grammar adapted for computers
Sets the tone for next years



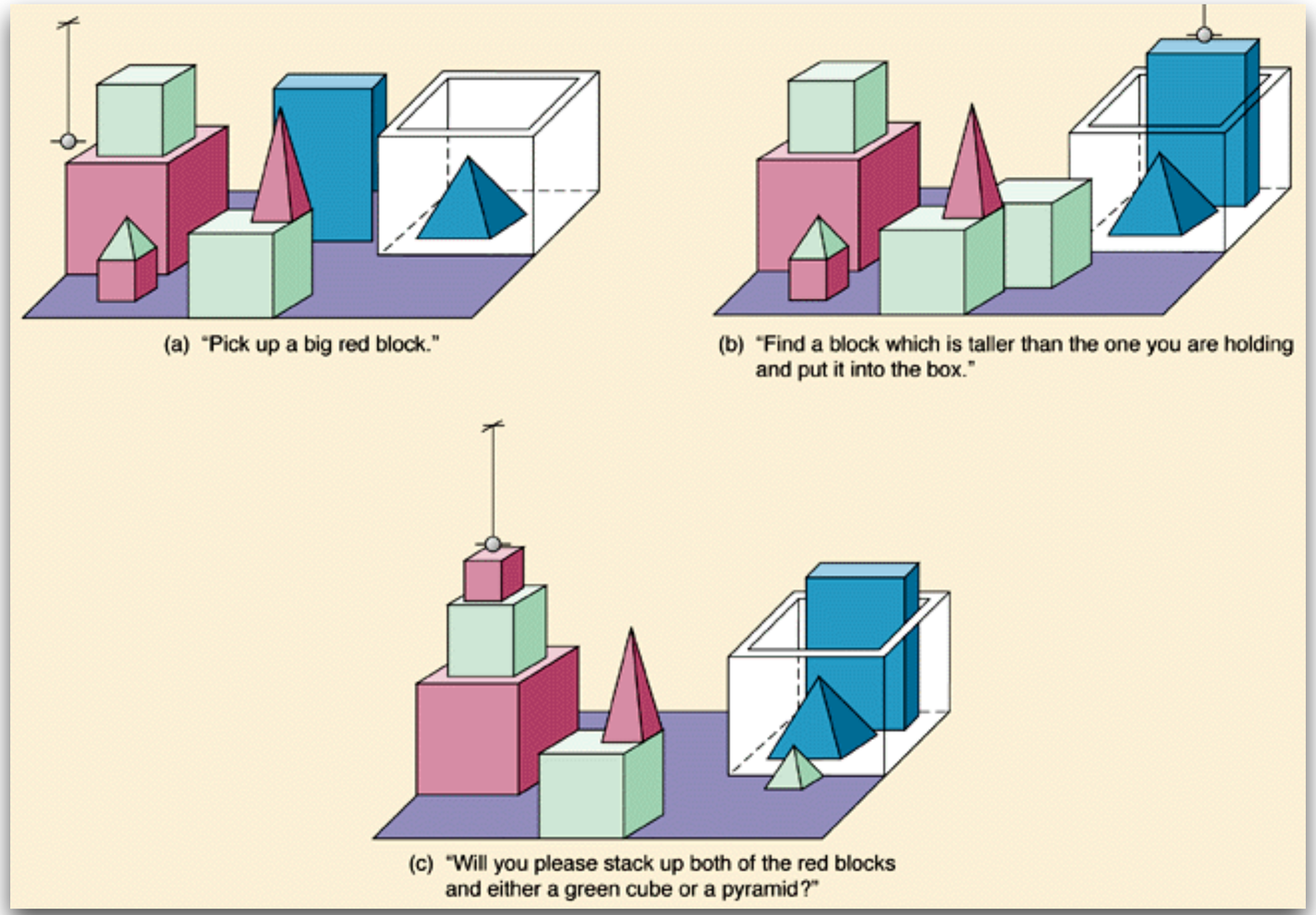
The things are not so good in the 1960s

- 1964 — Automatic Language Processing Advisory Committee, ALPAC
- 1966 — ALPAC puts a report that there is no Machine Translation now and won’t be in the nearest future

John Robinson Pierce, Head of ALPAC

1970, MIT, SHRDLU

- Terry Winograd
- Looked like a huge success



Hardcoded a lot

Not expandable

```
(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))))
```

Coming closer

- 1974 — effective back propagation algorithm
- 1980s — empirical revolution: more data and more power, statistical approach is taking back
- 1984 — big leaps in machine translation from IBM, statistics is rolling

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

Our days

Now, statistical methods have outperformed symbolical and are dominating the NLP industry

- 2011, Siri, voice recognition and generation on your phone
- 2013, word2vec, effective word vectorisation
- 2014, IBM watson takes on clinical patient history analysis
- 2016, Google Translate works with > 100 language pairs
- 2018, NLP as a service on Amazon AWS and Google Cloud
- 2018, Bot is talking to the customer support

NLP Applications

- Machine translation
- Information retrieval
- Text classification
- Chat bots
- QA systems

State of NLP

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?

Practice: Telegram Bot

Bot registration

- Install/register at Telegram
- Go to https://telegram.me/BotFather
- /help will help you
- /newbot
- Come up with a username
- Get the API Token

How to run a bot

- pip install pyTelegramBotAPI
- https://github.com/eternnoir/pyTelegramBotAPI

What can a bot do?

- React to messages
- React to commands
- Provide menus
- Handle different content types
- Send back different content types
- Store the state (that’s on you though)

Demo Time

Exercises

- Run a simple bot that gets user's username
- Get the sentiment of the message with https://text-processing.com/docs/sentiment.html
- Run some of your NN models as a reply to an image sent to the bot
- Find a dependency tree and some properties of words with spaCy