Introduction to Natural Language Processing

Part I: Overview

Henning Wachsmuth

https://ai.uni-hannover.de

Outline of the course

I. Overview
   - Introduction
   - Applications
   - Challenges
   - Approaches

II. Basics of Linguistics

III. NLP using Rules

IV. NLP using Lexicons

V. Basics of Empirical Methods

VI. NLP using Regular Expressions

VII. NLP using Context-Free Grammars

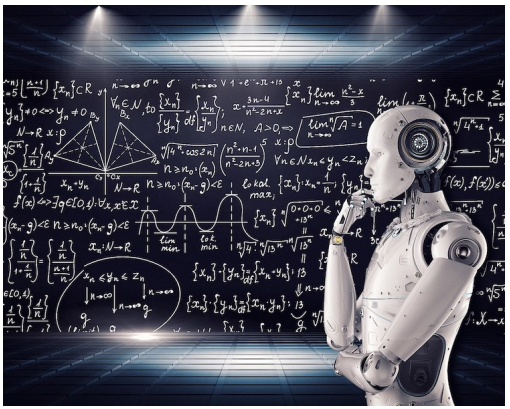
VIII. NLP using Language Models

IX. Practical Issues

Introduction

Natural Language Processing (NLP)

Natural language processing


https://wikimedia.org

- The study of computational methods for understanding and generating human-readable text (or speech)

  We mostly speak about text only in this course.
- The goal is to decode structured information from language, or to encode it in language.
- NLP is a subfield of AI, and one part of *computational linguistics*.

Computational linguistics

- Roughly, the intersection of computer science and linguistics
- Methods for tackling analysis and synthesis tasks from NLP
- Models to explain linguistic phenomena, using knowledge or statistics

Linguistics

- The study of natural language(s) in terms of form, meaning, and context

Natural Language Processing (NLP)

Analysis and Synthesis

Types of NLP tasks

- Analysis. The decoding of structured information from text
- Synthesis. The encoding of (structured) information into text

  Aka *natural language understanding (NLU)* and *natural language generation (NLG)*

Selected analysis tasks

- Token and sentence splitting
- Stemming and lemmatization
- Part-of-speech tagging
- Constituency/Dependency parsing
- Named/Numeric entity recognition
- Reference resolution
- Entity/Temporal relation extraction
- Topic/Sentiment/Spam classification
- Text scoring/grading

Selected synthesis tasks

- Lexicon creation
- Free text generation
- Sentence composition
- Discourse composition
- Spelling correction
- Summarization
- Text style transfer
- Cluster labeling

  ... among many other tasks

Natural Language Processing (NLP)

Example: Information Extraction

Task

- Identify entities, their attributes, and their relations in a given text
- Example. Extract company's founding dates from a news article

Output: Founded("Google", 1998)

Possible approach

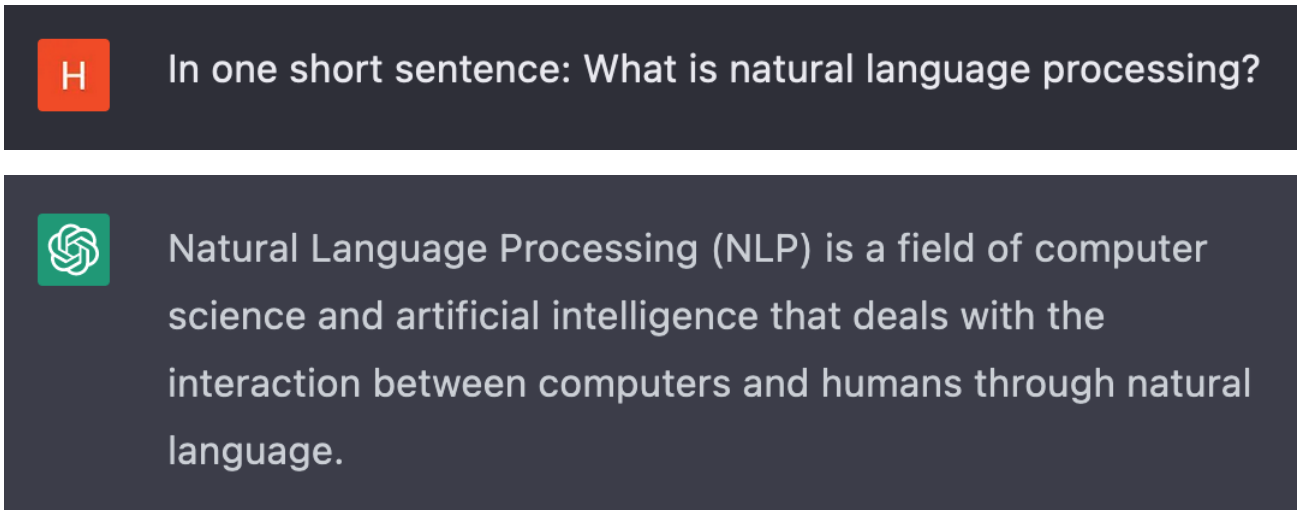
1. Lexical and syntactic preprocessing
2. Named and numeric entity recognition
3. Reference resolution
4. Entity relation extraction

Natural Language Processing (NLP)

Example: Language modeling

Task

- Extend a given text word by word until a suitable ending is reached.
- Example. Answer a user's question to a chatbot



Possible approach

1. Train general language model on huge amounts of text examples
2. Fine-tune model on question-answer training pairs

Natural Language Processing (NLP)

Terminology

Terms in NLP

- Task. A specific problem with a defined input and desired output
  Examples: Constituency parsing, summarization, ...
- Technique. A general way of how to analyze and/or synthesize a text
  Examples: Probabilistic parsing, language model, ...
- Algorithm. A specific implementation of a technique
  Examples: CKY parsing, GPT-3, ...
- Model. The configuration of an algorithm resulting from training
  Examples: CKY parsing on Penn Treebank, GPT-3 fine-tuned on a set of Q&A pairs, ...
- Approach. A computational method using model(s) to tackle a task
  Example: A method that fines phrases based on CYK parsing, ...
- Method. May refer to an algorithm, model, and/or approach
  Examples: As above
- Application. A technology that tackles a real-world problem using NLP
  Example: Watson, ChatGPT, ...

Applications

Applications

Applications

- Software that employs NLP to solve real-world problems
- This includes tools, systems, web services, and similar.

  The term *application* is also used in others ways in NLP.

Why applications?

- Automate human tasks and/or improve over human performance
- Use cases: Writing assistance, text analytics, conversational AI, etc.

Examples

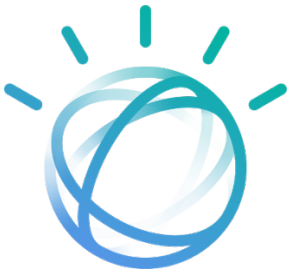
- Writing assistance. DeepL, Grammarly, Booking texts, Google Mail, ...
- Text analytics. IBM Watson, Facebook Ads & Targeting, Apple Mail, ...
- Conversational AI. ChatGPT, Google Assistant, Siri, Alexa, ...

Applications in this course

- The focus here is on computational methods rather than applications.
- Applications motivate why we deal with specific methods.

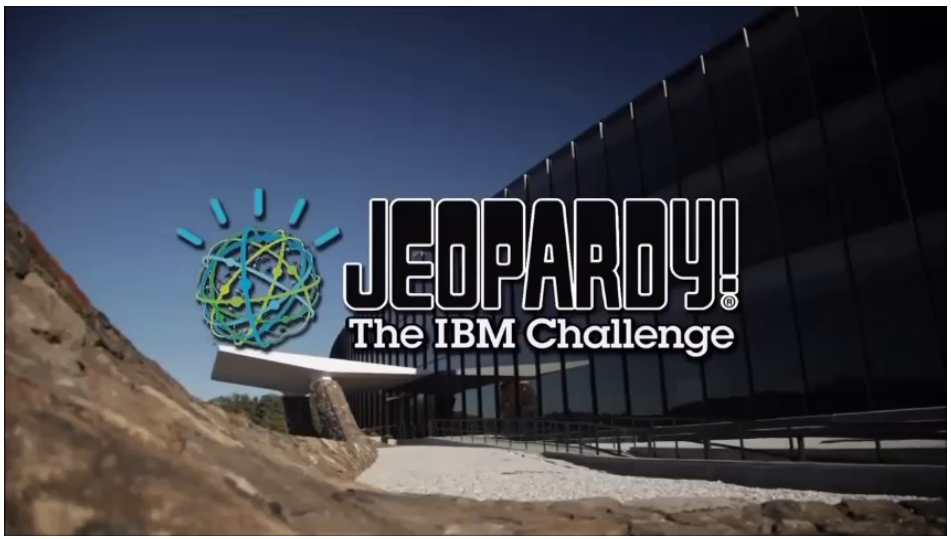
Example Application: Watson

IBM Watson



https://en.wikipedia.org

- A technology for text analytics and decision support
- Originally: A focused question answering system
- First showcase was the “Jeopardy!” task

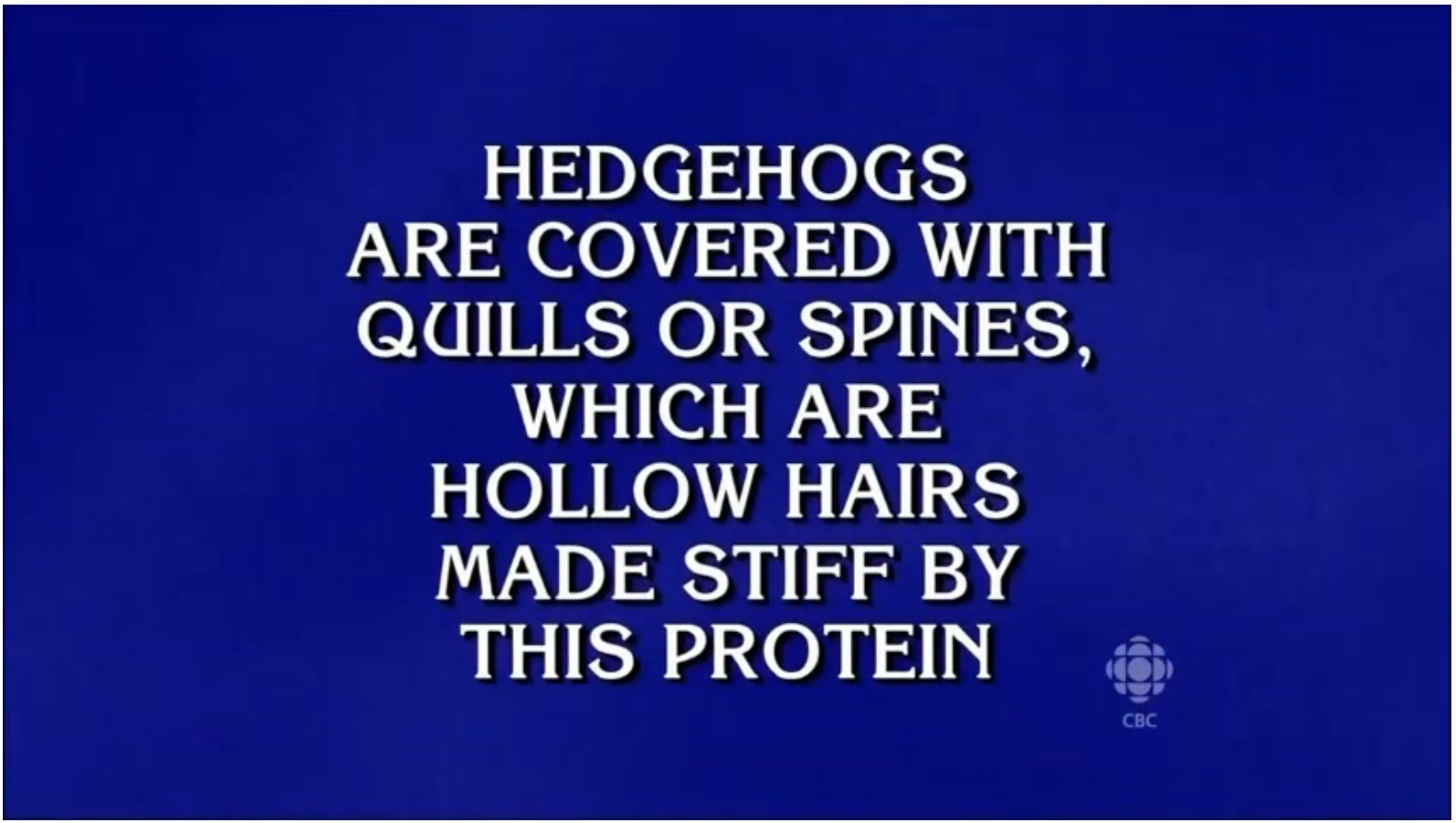


The IBM Challenge in 2011

- Watson plays against the best Jeopardy! champions
  `https://www.youtube.com/watch?v=P18EdAKuC1U`

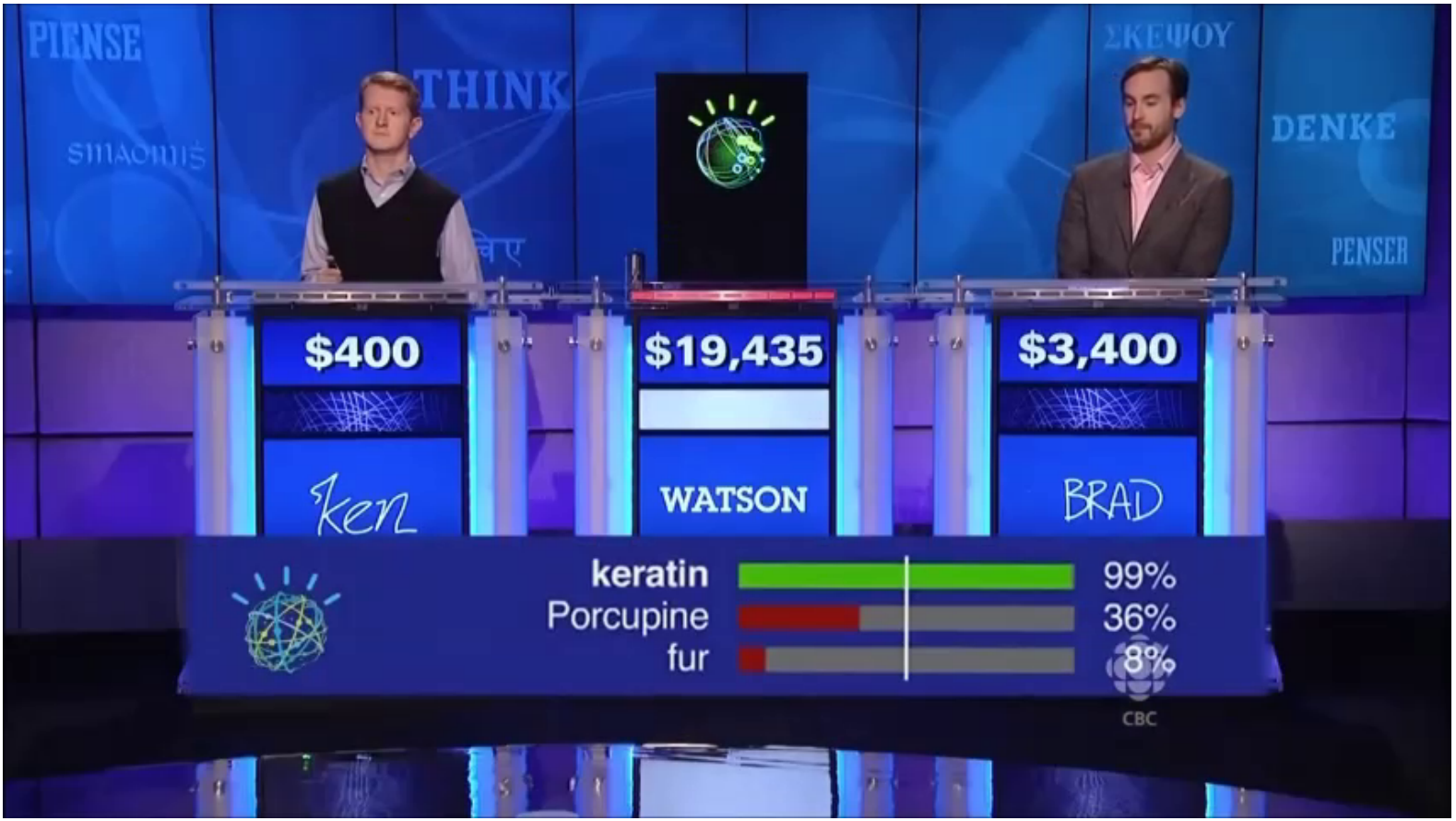
Example Application: Watson

Example “Question”



Example Application: Watson

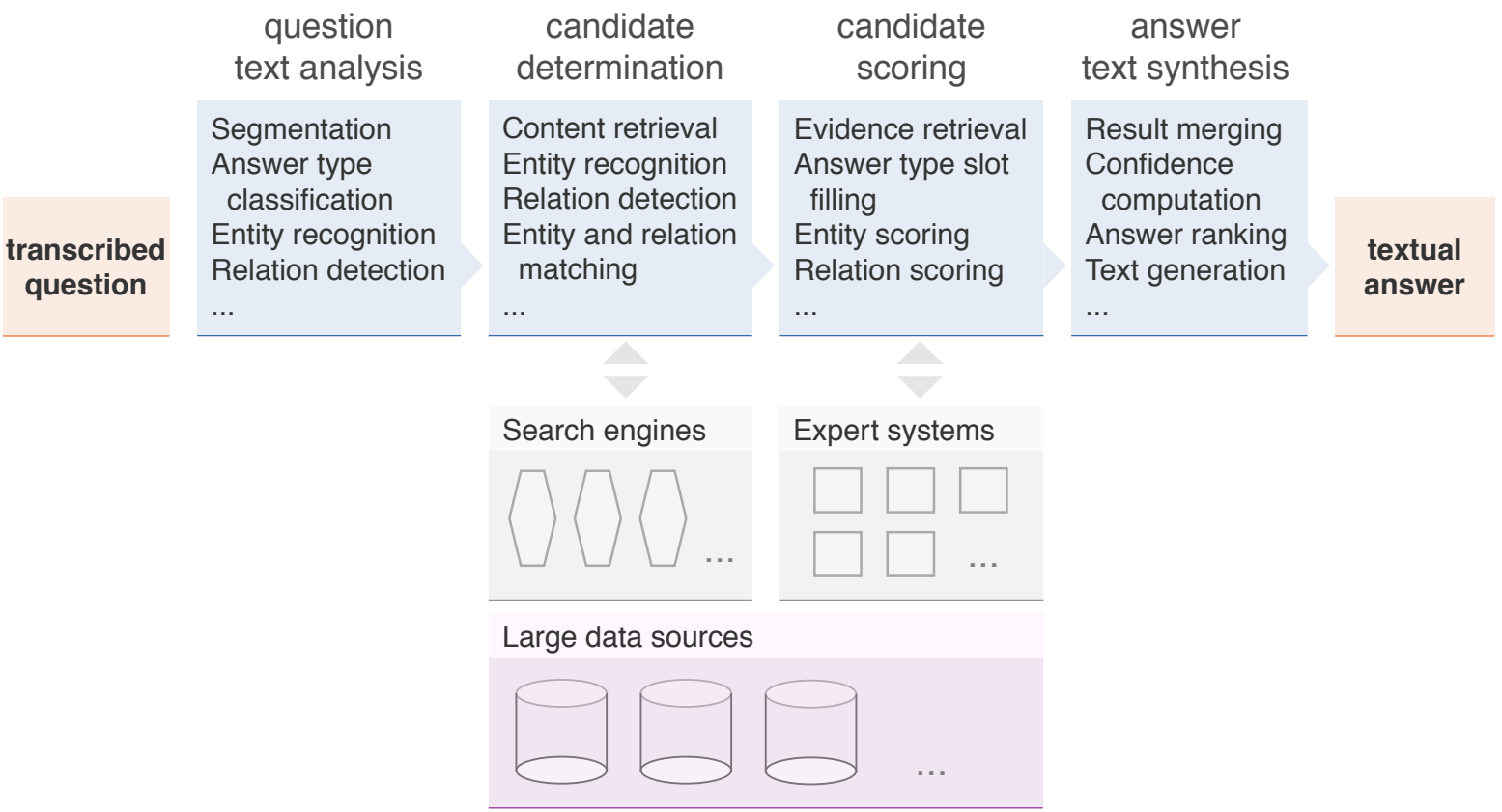
Watson's "Answer"



Example Application: Watson

NLP in Watson

Question answering process (simplified)



Applications

Evolution of NLP Applications

Selected milestones

- February 2011. Watson wins Jeopardy
  https://www.youtube.com/watch?v=P18EdAKuC1U

- October 2011. Siri starts on the iPhone
  https://www.youtube.com/watch?v=gUdVie_bRQo

- August 2014. Skype translates conversations in real time
  https://www.youtube.com/watch?v=RuAp92wW9bg

- May 2018. Google Duplex makes phone call appointments
  https://www.youtube.com/watch?v=pKVppdt_-B4

- February 2019. Project Debater competes in entire debates
  https://www.youtube.com/watch?v=nJXcFtY9cWY

- November 2022. ChatGPT leads conversations on any topic
  https://chat.openai.com

Observations

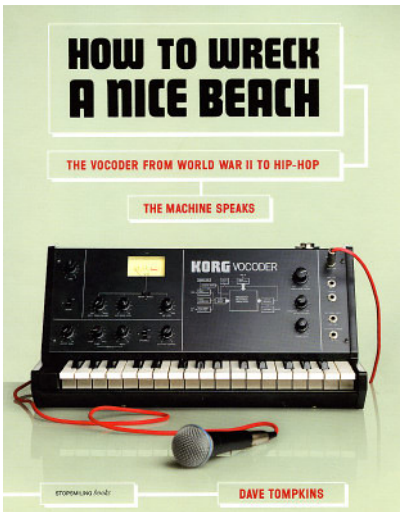
- NLP inside: All main analysis and synthesis tasks are tackled on text.
- None of these applications works perfectly.

Challenges

Challenges

Ambiguity

- Linguistic utterances allow for multiple interpretations.
- Fundamental challenge of processing natural language
- Pervasive across all language levels



https://flickr.com

Several types of ambiguity

- Phonetic. “wreck a nice beach”
- Word sense. “I went to the bank”.
- Part of speech. “I made her duck.”
- Attachment. “I saw a man with a telescope.”
- Scope. “I didn’t buy a car.”
- Coordination. “If you love money problems show up.”
- Speech act. “Have you emptied the dishwasher?”

Challenges

Limitations of Focus on Text

Purpose of "I never said she stole my money."

I never said she stole my money.	Someone else said it, but I didn't.
I *never* said she stole my money.	I simply didn't ever say it.
I never *said* she stole my money.	I might have implied it in some way. But I never explicitly said it.
I never said *she* stole my money.	I said someone took it. But I didn't say it was her.
I never said she *stole* my money.	I just said she probably borrowed it.
I never said she stole *my* money.	I said she stole someone else's money.
I never said she stole my *money*.	I said she stole something of mine. But not my money.

Challenges

Non-Standard Language

Colloquial language

- Non-standard writing. “@justinbieber Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever”
- Informal use. “This is sh*t” vs. “This is the sh*t”

Special phrases

- Tricky entities. “Let it Be was recorded”, “mutation of the for gene”, ...
- Idioms. “get cold feet”, “lose face”, ...
- Neologisms. “unfriend”, “retweet”, “hangry”, ...

Tricky segmentation

- Hyphens. “the New York-New Haven Railroad”
- Punctuation. “She was a Dr. I was not.”
- Whitespaces. “ 本を読む ”, “Just.Do.It.”

Challenges

Practical Issues

Common practical issues

- NLP faces effectiveness, efficiency, and robustness issues in practice.
- How to deal with such issues will be discussed at the end of this course.

Effectiveness issues

- Effectiveness. The extent to which the output of a method is correct
- Methods may not be effective enough for use in real-life applications.

Efficiency issues

- Efficiency. The run-time, space, or energy consumption of a method
- Methods may not be efficient enough when applied to big text amounts.

Robustness issues

- Robustness. The effectiveness of a method across domains of text
- Methods may not be robust enough on data different from training data.

Approaches

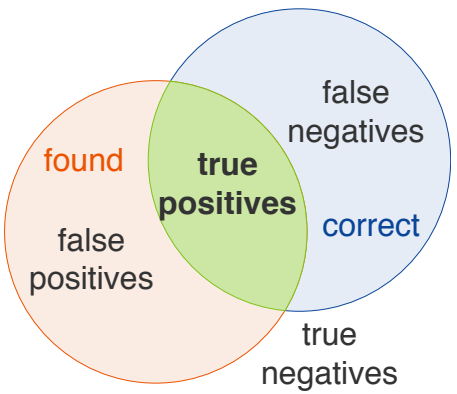
Approaches

Need for data

- NLP methods are meant to tackle specific analysis or synthesis tasks.
- To this end, they operationalize expert rules and/or statistical patterns.
- Rules and patterns are derived from analyses of training data.

Need for evaluation

- The output of NLP methods is rarely free of errors due to the outlined challenges.
- Thus, they are evaluated empirically on test data.
- The *effectiveness* of the methods is quantified in terms of metrics, such as accuracy.



Need for comparison

- It is unclear how good a measured effectiveness in a given task is.
- Approaches are thus compared to other methods, so called *baselines*.

Approaches

Text Corpora and Datasets

Text corpus

- A collection of real-world texts with known properties, compiled to study a language problem
- NLP methods are developed and tested on corpora.

Annotation

- An annotation marks a text or a span of text that represents an instance of a particular type of information.
- Annotations represent meta-information about the marked parts.
- The texts in a corpus are often annotated for the problem to be studied.

Dataset

- A sub-corpus used for developing and/or evaluating methods.
- Typical: a *training set* for development, a *validation set* for evaluation during development, and a *test set* for the final evaluation

Approaches

Inference Techniques

Knowledge-based inference

- Inference is based on manually-encoded expert knowledge.
- Knowledge is represented by rules, lexicons, grammars, and similar.

  We will see respective NLP techniques in the earlier part of this course.

Statistical inference

- Inference is based on statistical patterns found in training data.
- Patterns capture frequencies and/or manually-defined text features.

  We will see first respective NLP techniques in the later part of this course.

Neural inference

- Inference is based on statistical patterns found in training data.
- Patterns are automatically encoded in neural networks.

  Respective NLP techniques are treated in our master courses.

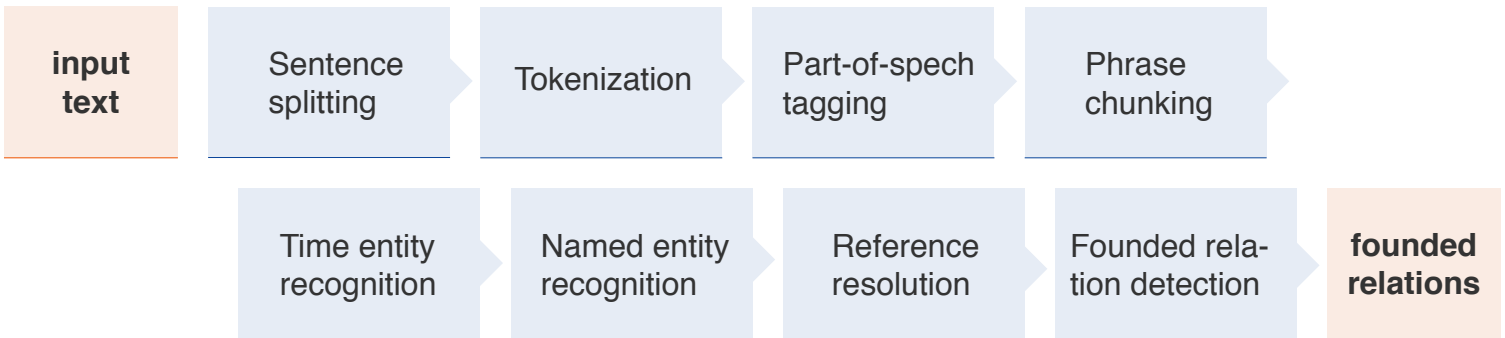
Approaches

Inference Processes

Pipeline approach

- The default way to realize an inference process is in form of a pipeline.
- A pipeline sequentially applies a set of NLP algorithms to a given text.

Example: Pipeline for founding date extraction



Alternatives

- Joint approach. Tackle multiple analysis/synthesis tasks simultaneously
- Neural approach. Operates on the raw input text (or tokens)

  Even with these, some kind of pipeline is used for most inference processes.

Approaches

Development and Evaluation

Input

- Task. An NLP task to be tackled
- Text corpus. A corpus, split into development and evaluation datasets

A typical development process

1. Analyze on training set how to best tackle the task.
2. Develop approach based on some technique that tackles the task.
3. Evaluate the effectiveness of the approach on the validation set.
4. Repeat steps 1–3 until effectiveness cannot be improved anymore.
5. Evaluate the effectiveness of the final approach on the test set.

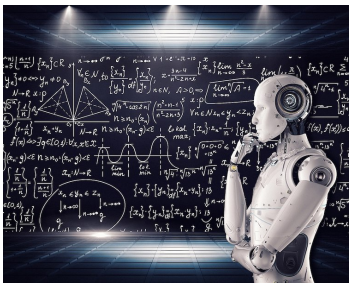
Output

- Approach. An NLP approach that tackles the given task
- Results. Empirical effectiveness measurements of the approach

Conclusion

Conclusion

Natural language processing

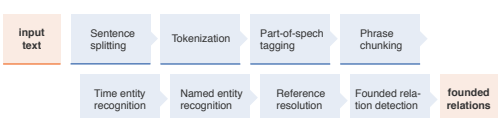


- Computational understanding and generation of text
- Analyses and syntheses at several language levels
- Disruptive applications such as conversational AI

Challenges

- Natural language is ambiguous in several ways
- Understanding requires context and world knowledge
- NLP aims to be effective, efficient, and robust

This course



- Teaches how to develop basic NLP methods
- Covers several tasks and techniques
- Covers design, implementation, and evaluation

References

Some content and examples taken from

- Emily M. Bender (2018). 100 Things You Always Wanted to Know about Semantics & Pragmatics But Were Afraid to Ask. Tutorial at the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), http://faculty.washington.edu/ebender/papers/Bender-ACL2018-tutorial.pdf.
- Daniel Jurafsky and Christopher D. Manning (2016). Natural Language Processing. Lecture slides from the Stanford Coursera course. https://web.stanford.edu/~jurafsky/NLPCourseraSlides.html.
- Matthias Hagen (2018). Natural Language Processing. Slides from the lecture at Martin-Luther-Universität Halle-Wittenberg. https://studip.uni-halle.de/dispatch.php/course/details/index/8b17eba74d69784964cdefc154bb8b95.
- Daniel Jurafsky and James H. Martin (2009). Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. Prentice-Hall, 2nd edition.
- Christopher D. Manning and Hinrich Schütze (1999). Foundations of Statistical Natural Language Processing. MIT Press.
- Henning Wachsmuth (2015): Text Analysis Pipelines — Towards Ad-hoc Large-scale Text Mining. LNCS 9383, Springer.