Statistical Natural Language Processing

Part I: Overview

Henning Wachsmuth

https://ai.uni-hannover.de

Outline of the Course

I. Overview
   - Introduction
   - Applications
   - Techniques
   - Conclusion

II. Basics of Data Science

III. Basics of Natural Language Processing

IV. Representation Learning

V. NLP using Clustering

VI. NLP using Classification and Regression

VII. NLP using Sequence Labeling

VIII. NLP using Neural Networks

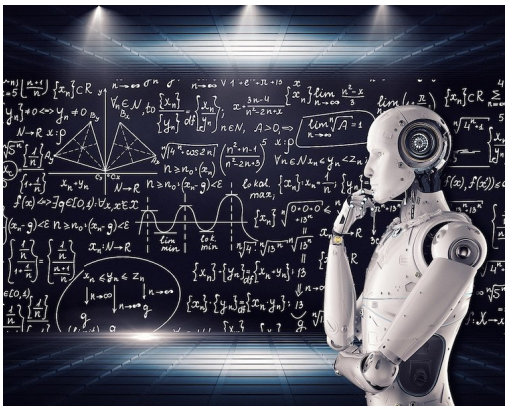
IX. NLP using Transformers

X. 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
- Technologies for natural language processing
- 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 inference of structured information from text (decoding)

  Analysis tasks are referred to as *natural language understanding (NLU).*
- Synthesis. The generation of text from structured information or from other text (encoding)

  Synthesis tasks are referred to as *natural language generation (NLG).*

Selected analysis tasks

- Token and sentence splitting
- Syntactic parsing
- Entity recognition
- Reference resolution
- Relation extraction
- Topic detection
- Sentiment analysis

Selected synthesis tasks

- Grammatical error correction
- Sentence generation
- Discourse composition
- Summarization
- Text style transfer
- Cluster labeling
- Lexicon creation

Natural Language Processing (NLP)

Development and Evaluation

Need for data

- NLP methods 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 ambiguity of language.
- Thus, they are evaluated empirically on test data.
- The *effectiveness* of methods is quantified with measures such as accuracy.

https://pixabay.com

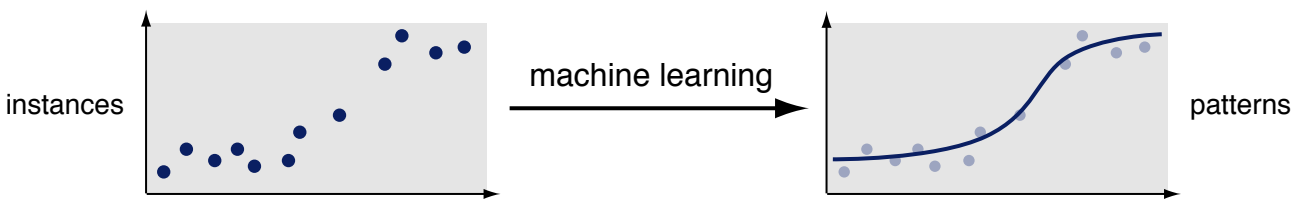
Need for comparison

- It is unclear per se how good a measured value is in a given task.
- Methods are thus compared to other methods, so called *baselines*.

Statistical Natural Language Processing (SNLP)

Statistical NLP

- Most recent NLP methods are data-driven, that is, they operationalize statistical patterns derived from training data.
- We will focus on such statistical (as opposed to rule-based) methods.
- The patterns are mostly found using *machine learning*.



Machine learning

- The study of algorithms that learn to tackle tasks from training data
- The algorithms generalize found statistical patterns into *models*.
- Models compare inputs, infer labels, generate text, or similar.
- Feature-based. Models learned from human-defined input features
- Neural. Models learned directly from the input using neural networks

Statistical Natural Language Processing (SNLP)

Rule-based vs. Statistical NLP

Rule-based methods

- Decision rules, regexes, ...
- Based on expert knowledge
- Each rule usually very precise

\+ Often match human intuition
\+ Behavior often easy to explain
\+ Simple tasks well-controllable
– Complex tasks hard to deal with
– Weighting hard to integrate
– Number of rules may explode
– Rules may be just hard to find

Statistical methods

- Classifiers, language models, ...
- Based on (often huge) data
- Pattern set aims to be effective

– Do not always match intuition
– Behavior often hard to explain
– Overhead for simple tasks
\+ Key to address complex tasks
\+ Weighting is a core concept
\+ Models may get very complex
\+ Data defines what can be found

Notice

- Due to their advantages, rule-based methods still remain in industry.
- Statistical methods often more effective; they build on rule-based ideas

Statistical Natural Language Processing (SNLP)

Terminology

Terms in SNLP

- Task. A specific problem with a defined input and desired output
  Examples: Classification of sentiment polarity, generation of a text summary, ...
- Technique. A general way of how to analyze and/or synthesize a text
  Examples: Feature-based clustering, transformer-based text generation, ...
- Algorithm. A specific implementation of a technique
  Examples: k-means, BART, ...
- Model. The configuration of an algorithm resulting from training
  Examples: 5-means trained on data xxx, BART fine-tuned on data yyy
- Approach. A computational method using model(s) to tackle a task
  Example: A method that summarizes argumentative text using fine-tuned BART
- Application. A technology that tackles a real-world problem using NLP
  Example: Google Assistant

Notice

- Informally, the terms method, algorithm, model, and approach are often used more or less interchangeably.

Applications

Applications

Term “application” in NLP

- Approaches. Developed approaches process new data
- Downstream tasks. General NLP techniques are used for specific tasks

  Example: Sentiment analysis applies techniques such as text classification.
- Technologies. Developed approaches are deployed in software.

  This is the kind of applications that meant here.

Application in technologies

- Software that uses NLP to solve real-world problems
- This includes tools, systems, web services, and similar.
- Main application areas include human-machine interaction, automation, and data science.

  Examples follow on the next slides

https://de.wikipedia.org

Applications in this course

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

Applications

NLP in End-User Applications

Human-machine interaction (HCI)

- Chatbots. ChatGPT, WHO Coronavirus Bot, HelloFresh Freddy, ...
- Intelligent assistants. Google Assistant, Siri, Alexa, Cortana, ...
- Faceted review search. Booking, TripAdvisor, ...

Automation

- Writing assistance. Grammarly, LanguageTool, Textio, ...
- Information extraction. Various mail tools, web pages, ...
- Machine translation. DeepL, Google Translate, Skype live translation, ...

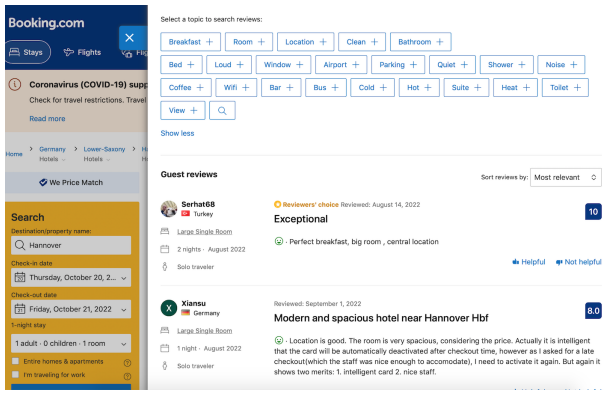
Data Science

- Text analytics. IBM Watson, Microsoft Azure, Amazon Web Services, ...
- Knowledge bases. DBpedia, Google Knowledge Graph, ...
- Advertising. Facebook Ads & Targeting, Adobe Advertising Cloud, ...

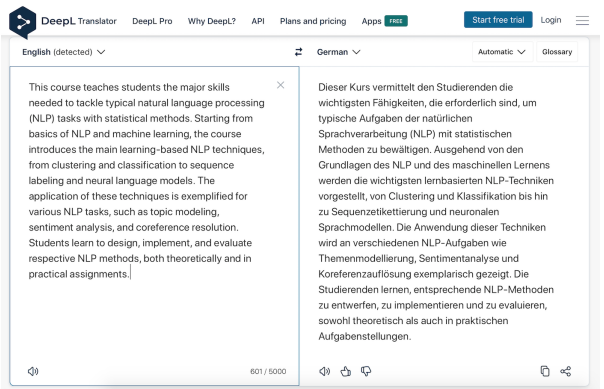
Applications

Selected Applications

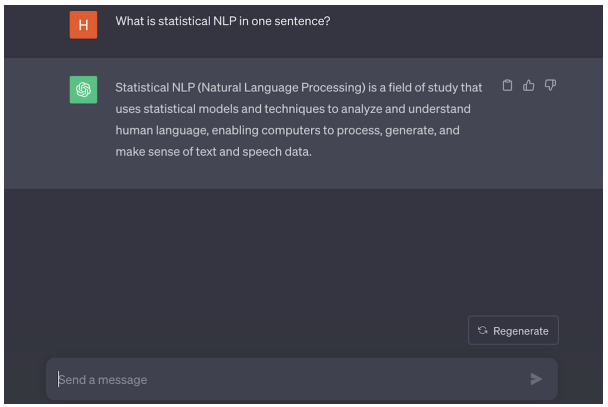
Booking booking.com



DeepL deepl.com



ChatGPT thttps://chat.openai.com

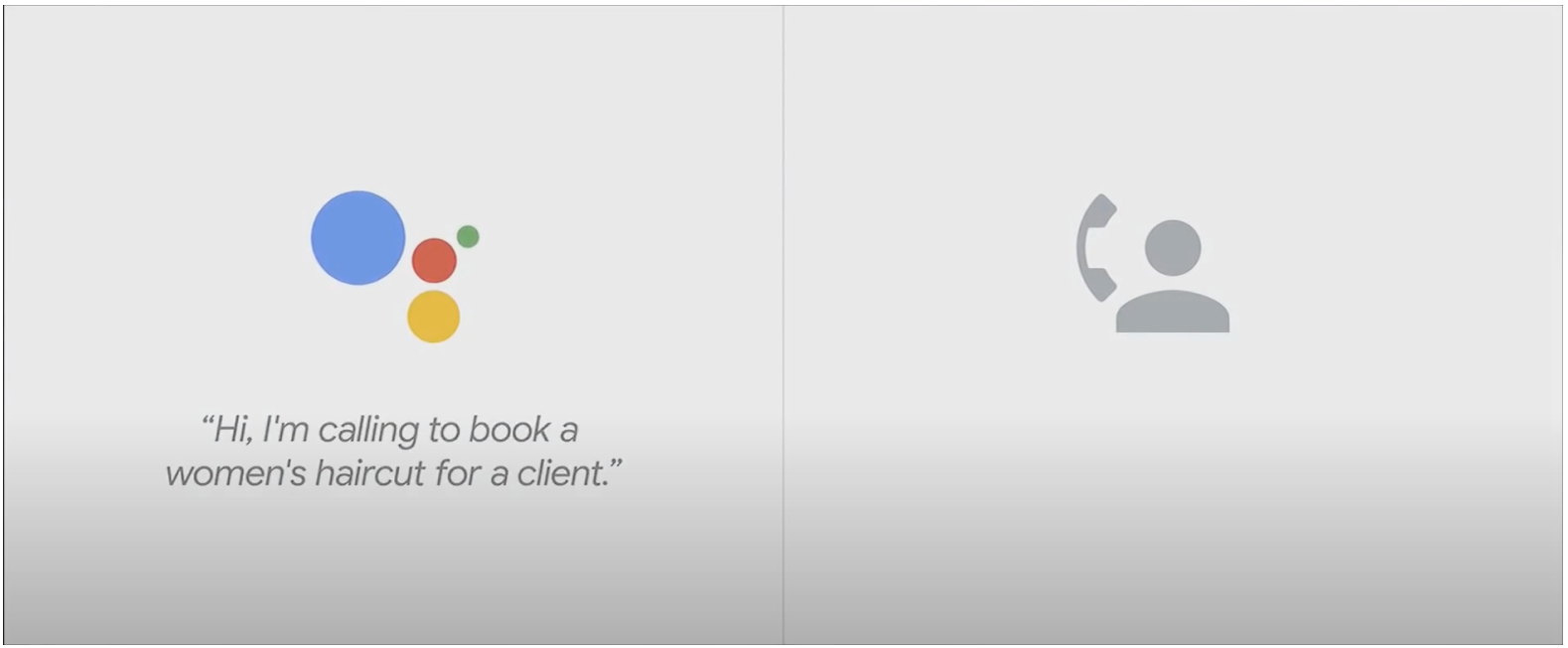


Google Assistant youtube.com

Example Application

Google Assistant making a phone call

- The technology behind is called Google Duplex

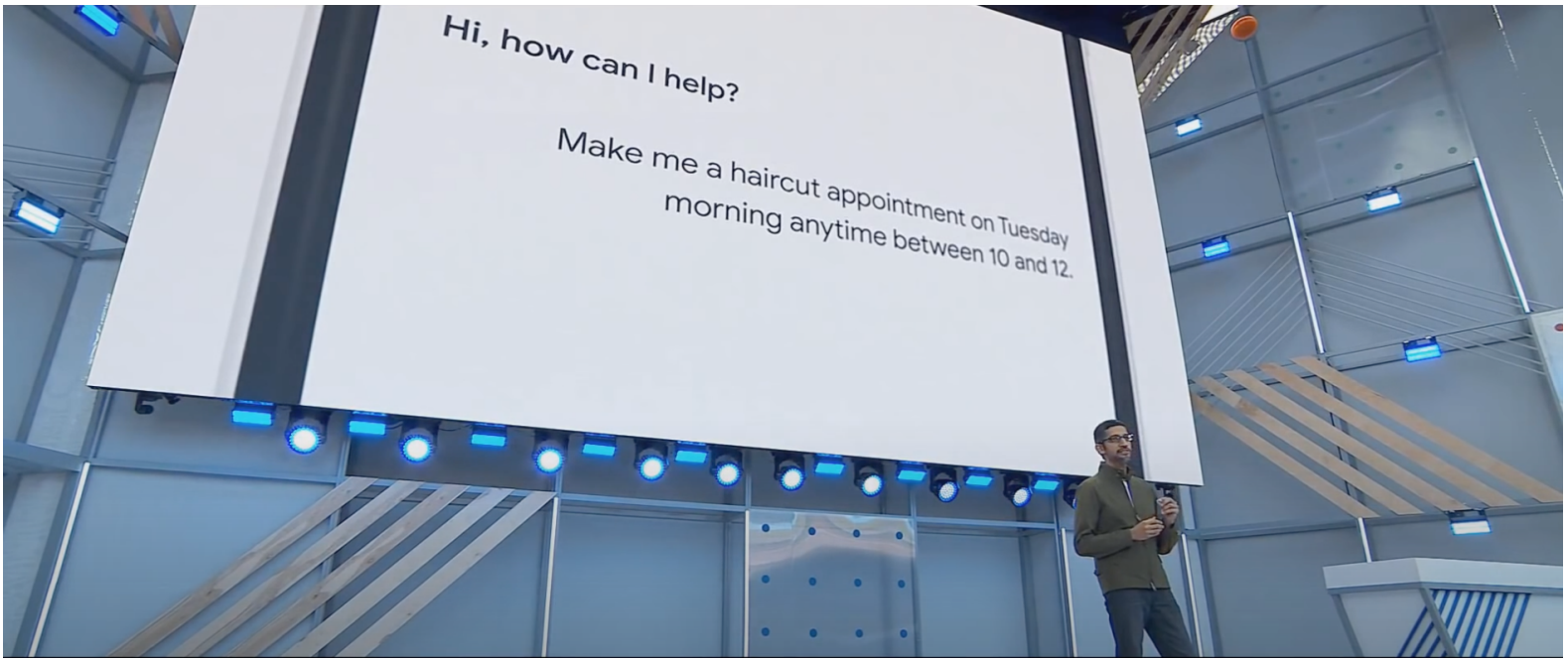


https://www.youtube.com/watch?v=pKVppdt_-B4

(screenshots on this and the following slides are taken from this video)

Example Application

Google Duplex: Analyzing a Request

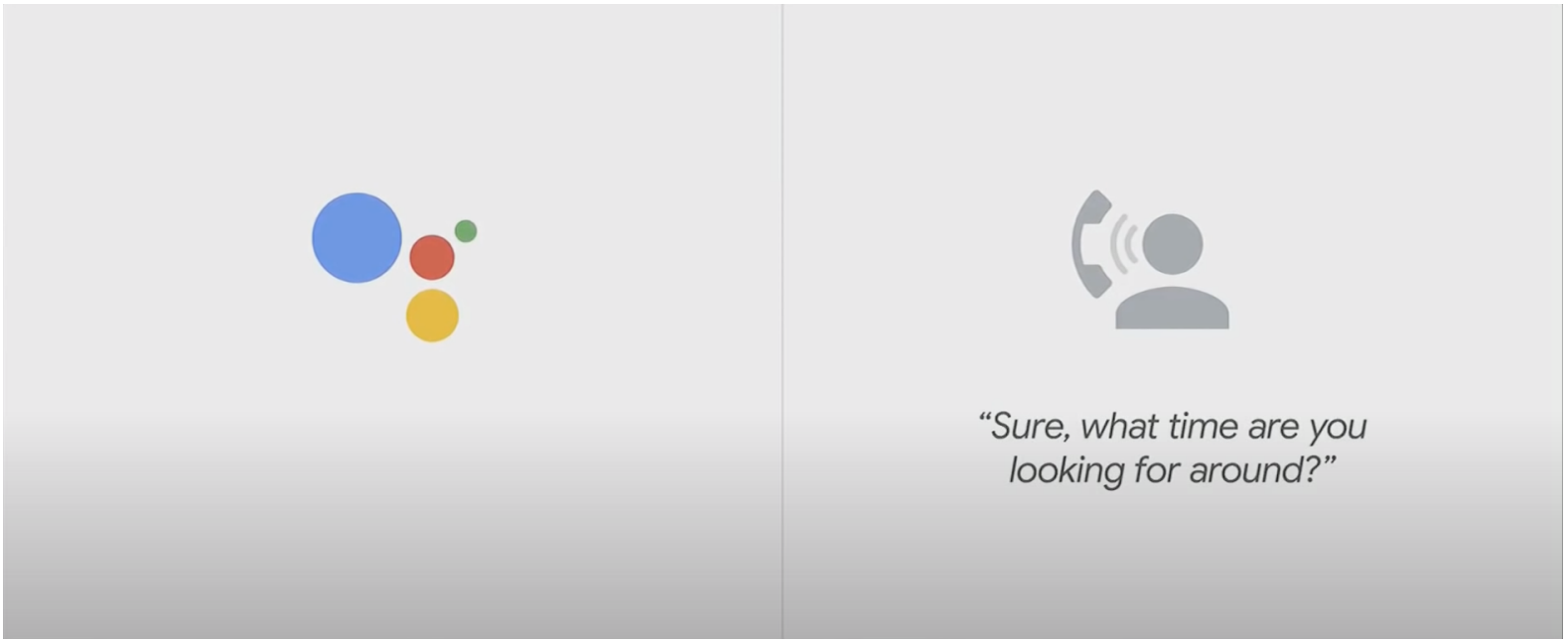


Conceptual steps

- Classify that this is a request to call a hairdresser.
- Extract service, date, and time period.
- Resolve these entities based on context information.

Example Application

Google Duplex: Analyzing a Question

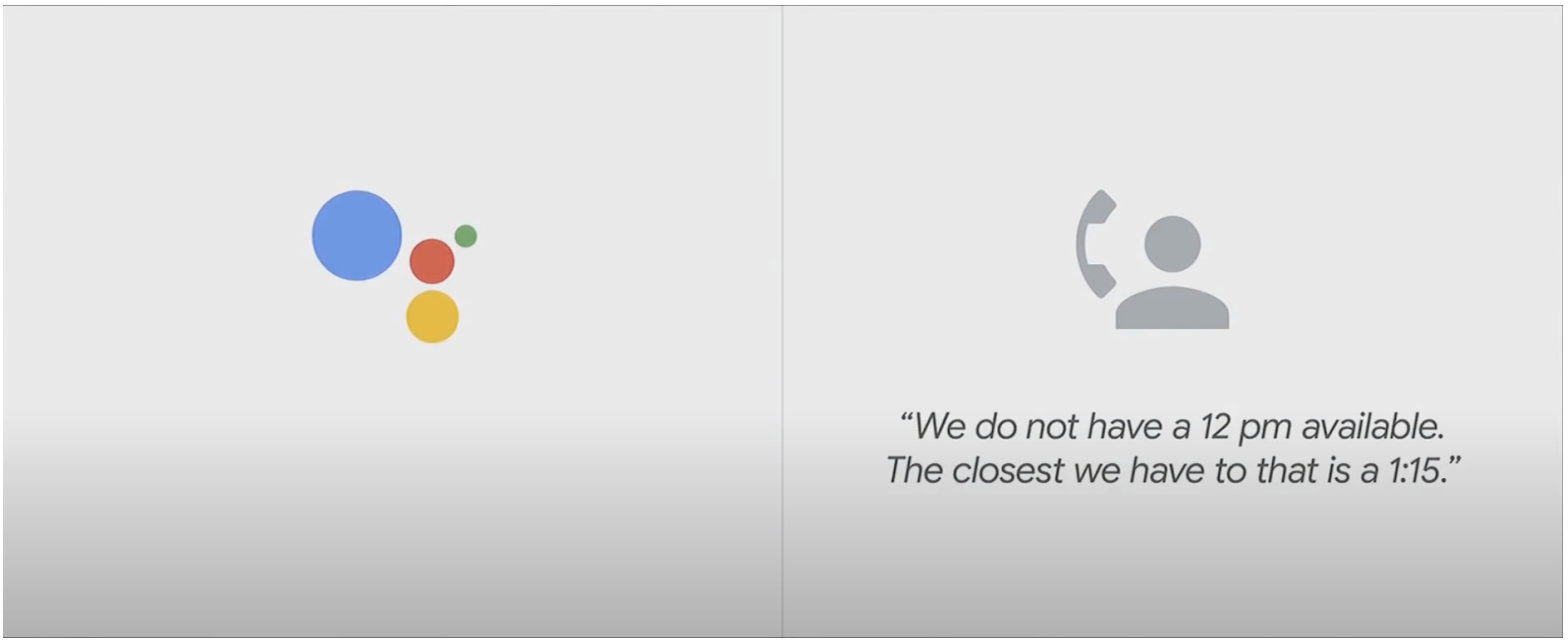


Conceptual steps

- Classify that the question asks for a point in time.
- Match with time extracted previously.

Example Application

Google Duplex: Analyzing Provided Information

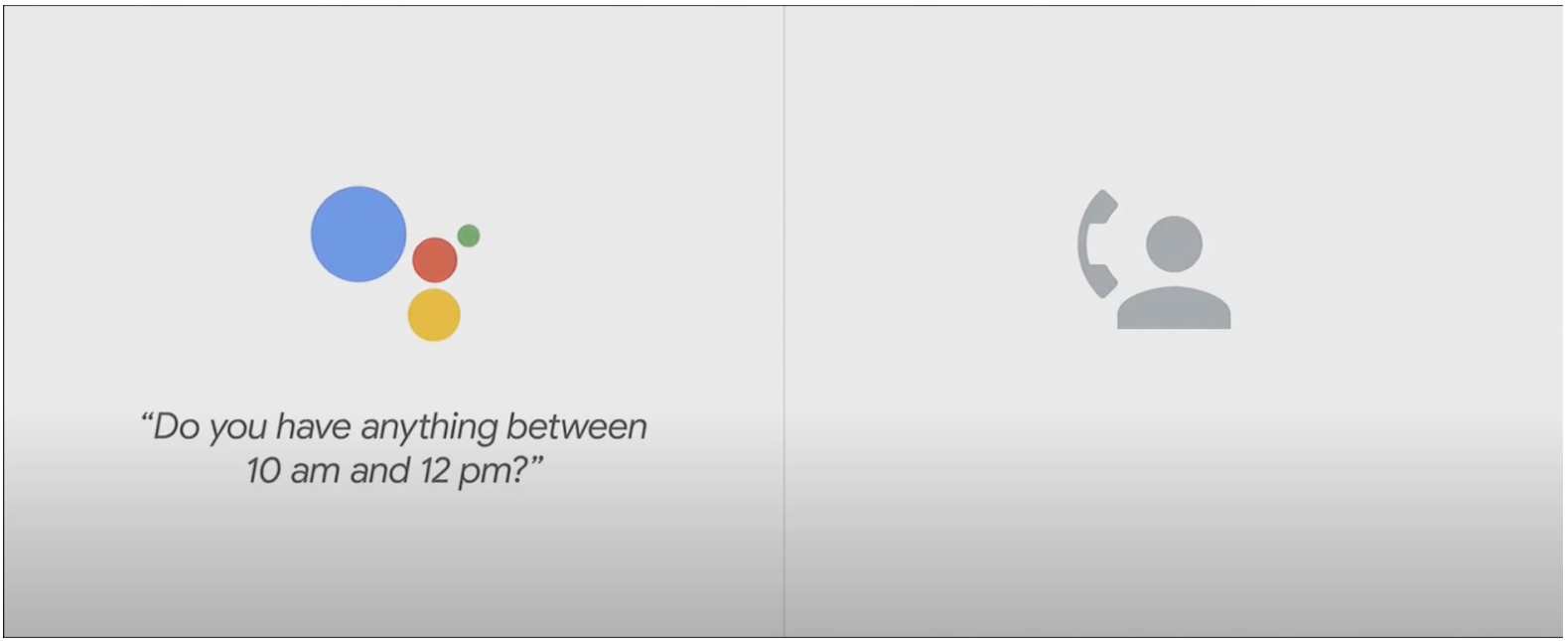


Conceptual steps

- Classify that this is a negative answer on the time.
- Classify that an alternative is provided.
- Extract time from alternative and match with previously extracted time.

Example Application

Google Duplex: Synthesizing a Request

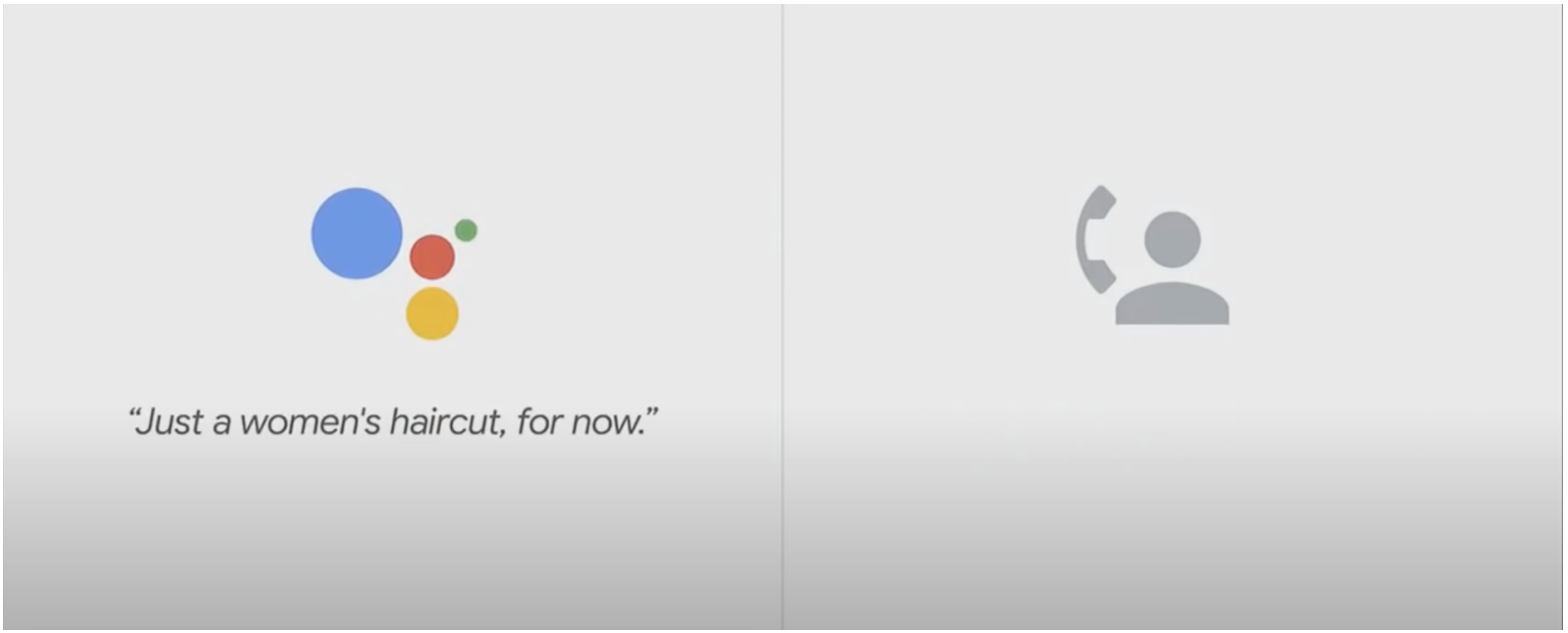


Conceptual steps

- Retrieve sentence template that fits information extracted previously.
- Fill template with information.

Example Application

Google Duplex: Synthesizing a Response

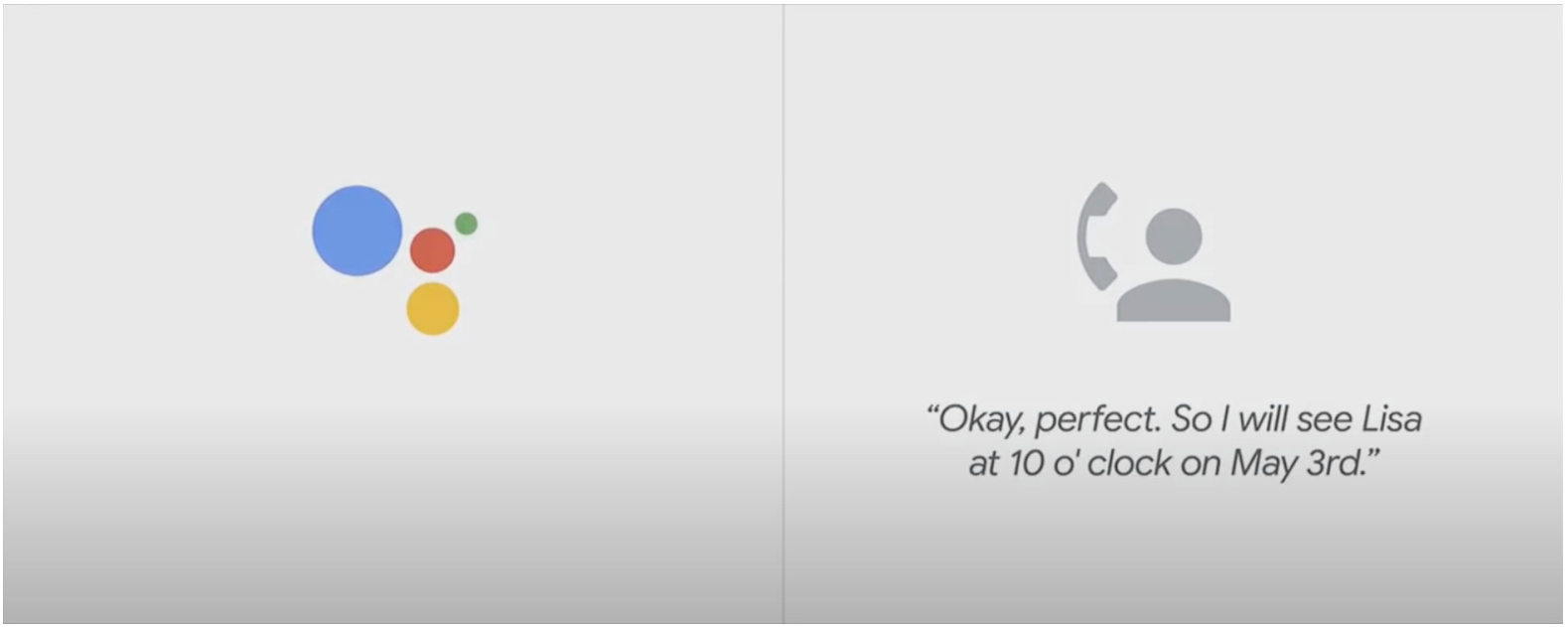


Conceptual steps

- Generate output sequence coherent to the question asked before.
- Ensure that the output sequence contains the information asked for.

Example Application

Google Duplex: Understanding that the Task is Done



Conceptual steps

- Extract all information provided by the dialogue partner.
- Match with predefined information needed to be found.
- Classify that this concludes the dialogue.

Techniques

Techniques in SNLP

Techniques

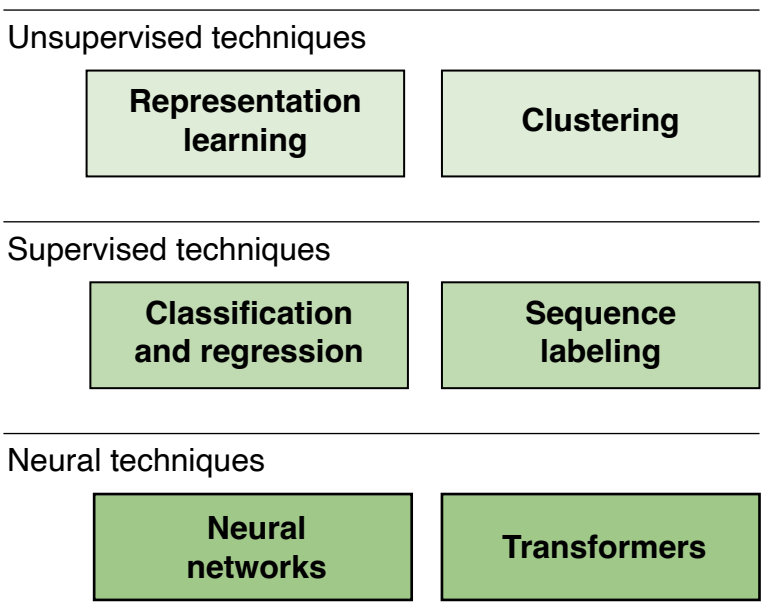
- General ways of how to analyze and/or synthesize some input text
- A technique defines a family of methods that follow the same principles.
- This course is structure according the six techniques considered.

Considered techniques

- Fully unsupervised techniques
- Fully supervised techniques
- Neural techniques, combining ideas from both

Practical Issues

- We will also look at issues to be dealt with when applying the techniques in practice



Techniques in SNLP

Properties of Techniques

Desired output

- Clustering. A set of instances is grouped into not-predefined classes.
- Classification. Each instance is assigned a predefined class label.
- Regression. Each instance is assigned a numeric value.
- Generation. Each instance is a linguistic utterance.

Interdependencies of the input

- Independent. Each instance is treated in isolation.
- Sequential. Instances are treated based on others in a sequence.
- Hierarchical. Instances are recursively decomposed into sub-instances.
- Fully dependent. All instances are treated jointly.

Learning process

- Unsupervised. No class labels/values used in development
- Supervised. Correct labels/values of training instances available

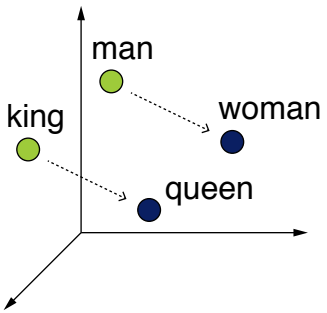
... and some others

Techniques in SNLP

Unsupervised Techniques

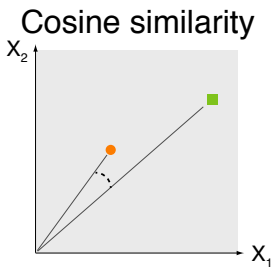
Representation learning

- Methods for formally representing textual instances
- Mostly, instances are mapped to real-valued vectors.
- Examples. Feature vectors, dense embeddings



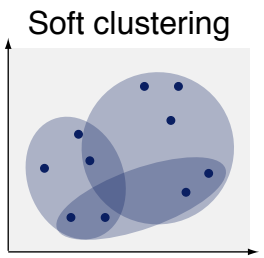
Similarity measures

- Methods that quantify how similar two instances are
- Mostly, instances are represented in vector form.
- Examples. Cosine score, euclidean distance



Clustering

- Methods that group a set of instances into $k \geq 1$ clusters
- Learned models group based on similarity measures.
- Examples. k-means, latent Dirichlet allocation

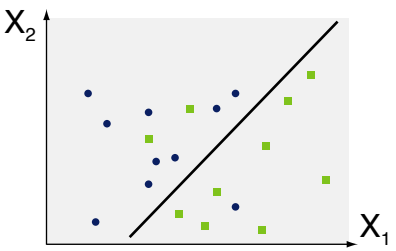


Techniques in SNLP

Supervised Techniques

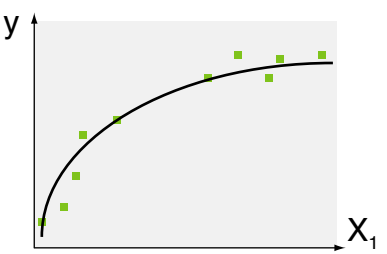
Classification

- Methods that assign one of $k > 1$ labels to an instance
- Used to categorize texts, spans, and relations
- Examples. Support vector machine, Naïve Bayes



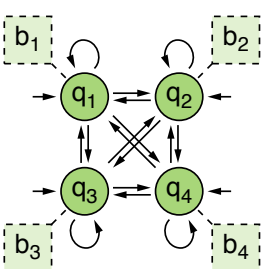
Regression

- Methods that assign a real value to an instance
- Used to scores texts or spans, or quantify likelihood
- Examples. Linear regression, SVR regression



Sequence labeling

- Methods that assign labels to consecutive instances
- Used to classify each span in a sequence of spans
- Examples. Hidden Markov model, cond. random field

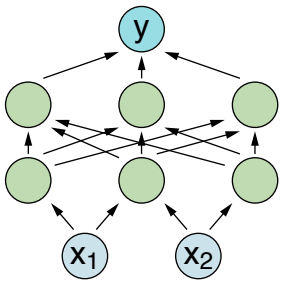


Techniques in SNLP

Neural Techniques

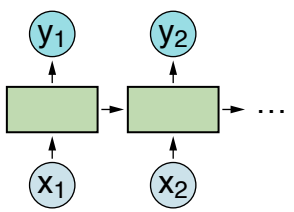
Feed-forward neural network

- Networks that predicts output values for a single instance
- Used for classification and regression tasks
- Examples. Multilayer perceptron, convolutional network



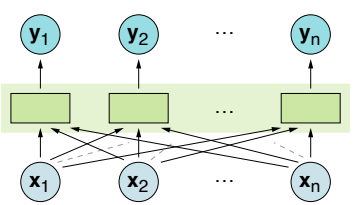
Recurrent neural network

- Networks that predict values for sequences of instances
- Used for sequence labeling and text generation tasks
- Examples. Bi-directional LSTMs, Seq2Seq models



Transformers

- Networks that predict values for sequences in parallel
- Used for any of the above analysis and synthesis tasks
- Examples. Left-to-right, bidirectional, encoder-decoder



Practical Issues

Common issues of NLP in practice

- NLP faces different effectiveness and efficiency issues in practice.
- Besides, ethical issues may come up when applying NLP
- How to deal with such issues will be discussed at the end of this course.

Effectiveness issues

- Methods may not be reliable enough for real-life applications.
- Methods may not work robustly on data different from the training data.

Efficiency issues

- Methods may not be efficient enough for real-life applications.
- The energy need of training may be prroblematic for the environment.

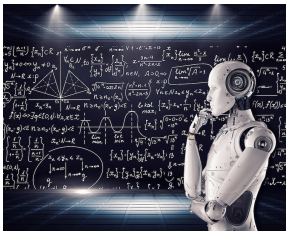
Ethical issues

- Methods may be biased against certain social groups.
- Methods may be employed in doubtful applications.

Conclusion

Conclusion

Statistical NLP

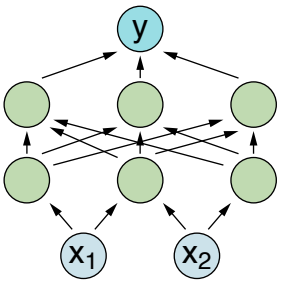


- Computational analysis and synthesis of text
- Development and evaluation based on textual data
- Statistical methods mostly rely on machine learning

Applications

- NLP used for HCI, automation, data science, and more
- All big tech companies employ NLP methods nowadays
- Examples include ChatGPT and Google Assistant

Techniques



- Various ways to group, classify, and score instances
- From simple similarity measures to neural transformers
- Effectiveness, efficiency, and ethics issues in practice