Introduction, administrivia

Statistical Methods in NLP 2
ISCL-BA-08

Çağrı Çöltekin
/tʃaːɾˈɯ tʃœltecˈɪn/
ccoltekin@sfs.uni-tuebingen.de

University of Tübingen
Seminar für Sprachwissenschaft

Summer Semester 2025

version: 0e94597+ @2025-04-25

What is this course about?

- This is the second course in the two-course series on *Natural Language Processing*
- A course focuses on machine learning techniques for working with linguistic data
- Focus on applications of the models to real tasks/problems

Prerequisites:

- ISCL-BA-06, ISCL-BA-07

Module: ISCL-BA-08

Text (sequence) classification

- Text classification is a very common NLP task
- Given a text we often want to assign one or more labels
- Assignment of the label(s) requires some level of 'language understanding'

Text classification: some examples

is it spam?

From: Dr Pius Ayim <mikeabass15@gmail.com>
Subject: Dear Friend / Lets work together

Dear Friend,

My name is Dr. Pius Anyim, former senate president of the Republic Nigeria under regime of Jonathan Good-luck.

I am sorry to invade your privacy; but the ongoing ANTI-CORRUPTION GRAFT agenda of the rulling government is a BIG problem that I had to get your contact via a generic search on internet as a result of looking for a reliable person that will help me to retrieve funds I deposited at a financial institute in Europe.

…

* From my 'spambox' which I stopped checking regularly long time ago.

Text classification: some examples

is the customer happy?

```
I never understood what's the BIG deal behind this album. Yes,
the production is wonderfull but the songwriting is childish and
rubbish. They definitly can not write great lyrics like Bob
Dylan sometimes do. "God Only Know" and "Wouldnt Be nice" are
indeed masterpieces...but the rest of the album is background
music.
```

Text classification: some examples

is the customer happy?

> I never understood what's the BIG deal behind this album. Yes, the production is wonderfull but the songwriting is childish and rubbish. They definitly can not write great lyrics like Bob Dylan sometimes do. "God Only Know" and "Wouldnt Be nice" are indeed masterpieces...but the rest of the album is background music.

> @DB_Bahn mußten sie für den Sauna-Besuch zuzahlen ?

Text classification: some examples

is the customer happy?

```
I never understood what's the BIG deal behind this album. Yes,
the production is wonderfull but the songwriting is childish and
rubbish. They definitly can not write great lyrics like Bob
Dylan sometimes do. "God Only Know" and "Wouldnt Be nice" are
indeed masterpieces...but the rest of the album is background
music.
```

```
@DB_Bahn mußten sie für den Sauna-Besuch zuzahlen ?
```

- Sentiment analysis is one of the popular applications of text classification

Text classification: some examples

which language is this text in?

```
Član 3. Svako ima pravo na život, slobodu i ličnu bezbjednost.
```

Text classification: some examples

which language is this text in?

```
Član 3. Svako ima pravo na život, slobodu i ličnu bezbjednost.
```

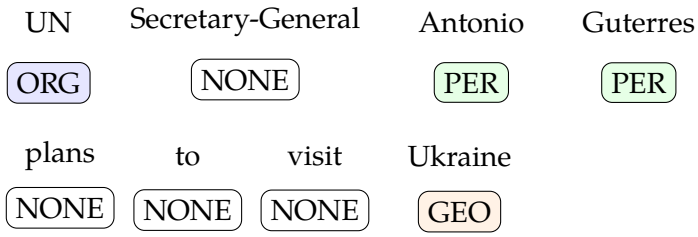
- Detecting language of the text is often the first step for many NLP applications.
- Easy for the most part, but tricky for
  - closely related languages
  - text with code-switching

Text classification: More examples

- Who wrote the book?
- Find the author's
  - age
  - gender
  - political party affiliation
  - native language
- Is the author/speaker depressed?
- What is the proficiency level of a language learner?
- What grade should a student essay get?
- What is the diagnosis, given a doctor's report?
- What category should a product be listed based on its description?
- What is the genre of the book?
- Which department should answer the support email?
- Is this news about
  - politics
  - sports
  - travel
  - economy
- Is the web site an institutional or personal web page?

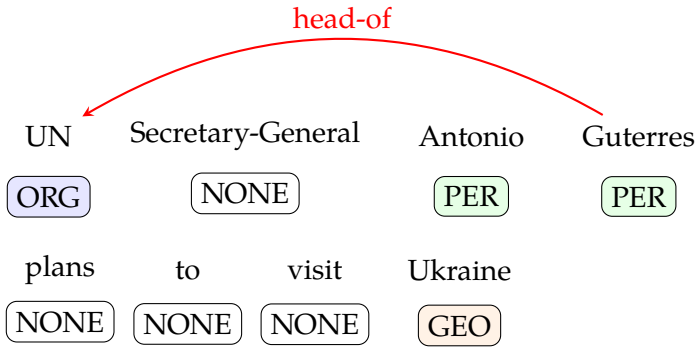
Sequence labeling

Example: entity recognition



- Typical entities of interest include: people, organizations, locations
- Can be application specific, e.g., drug/disease names, chemical components, legal entities
- Many NLP problems can be cast as sequence labeling problems: extractive summarization, question answering, ...

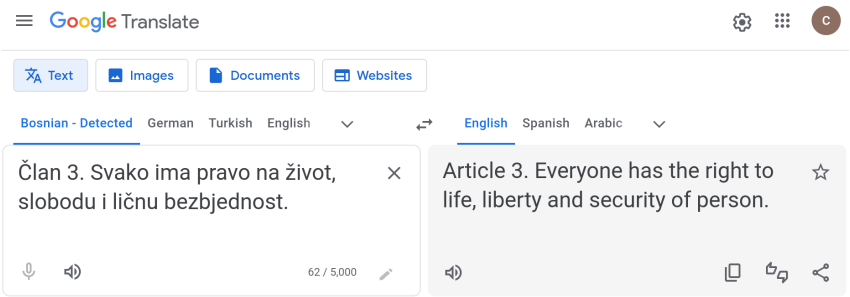
Relation extraction



- For many other tasks, we do not only need entities, but the relations between them
- Other similar applications include: dependency parsing, semantic role labeling, ...

Text generation

Example: machine translation



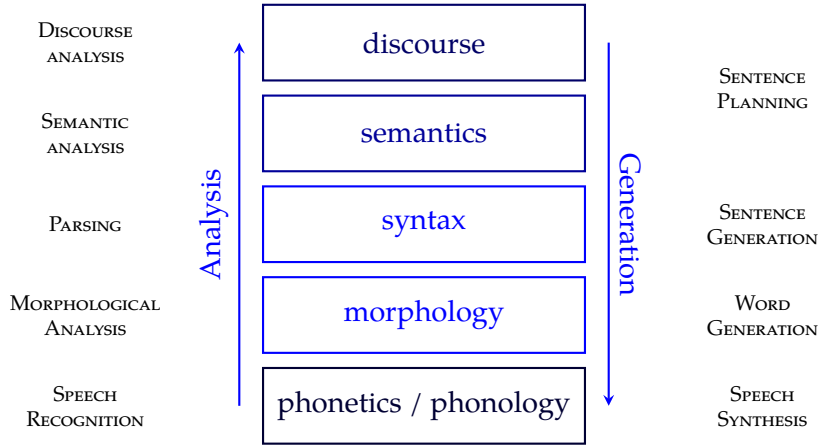
Text generation

Other examples

- Summarization
- Question answering
- Caption generaion
- Data to text generation

How do we solve these problems?

a historical perspective



How do we solve these problems?

more recently

On the word 'statistical'

> *But it must be recognized that the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term. — Chomsky (1968)*

- Some linguistic traditions emphasize(d) the use of 'symbolic', rule-based methods
- Some NLP systems are based on rule-based systems (esp. from 80's 90's)
- Virtually, all modern NLP systems are statistical

What is difficult with NLP?

- Combinatorial problems - computational complexity
- Ambiguity
- Data sparseness

NLP and computational complexity

- How many possible parses a sentence may have?
- How many ways can you align two (parallel) sentences?
- How many operations are needed for calculating probability of a sentence from the probabilities of words in it?

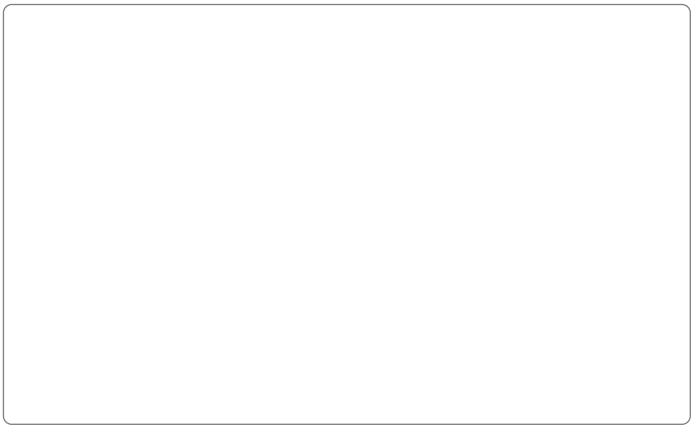
NLP and computational complexity

- How many possible parses a sentence may have?
- How many ways can you align two (parallel) sentences?
- How many operations are needed for calculating probability of a sentence from the probabilities of words in it?
- Many similar questions we deal with have an exponential search space
- Naive approaches often are computationally intractable

Combinatorial problems

A typical linguistic problem: parsing

How many different binary trees can span a sentence of N words?



Combinatorial problems

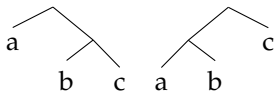
A typical linguistic problem: parsing

How many different binary trees can span a sentence of N words?

Combinatorial problems

A typical linguistic problem: parsing

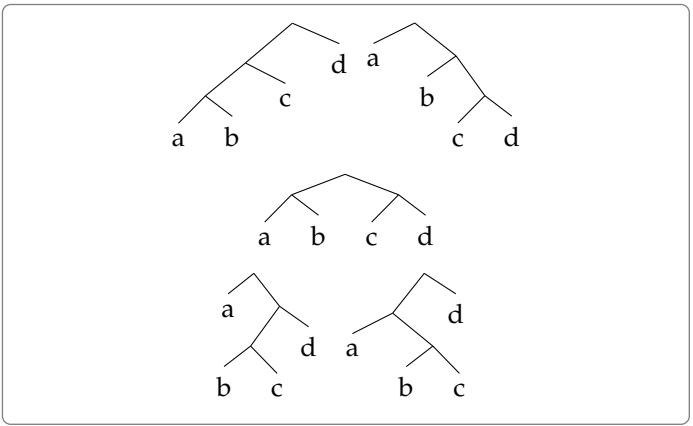
How many different binary trees can span a sentence of N words?



Combinatorial problems

A typical linguistic problem: parsing

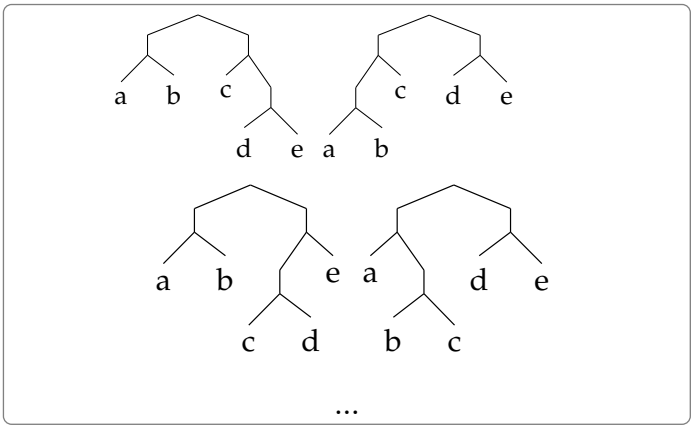
How many different binary trees can span a sentence of N words?



Combinatorial problems

A typical linguistic problem: parsing

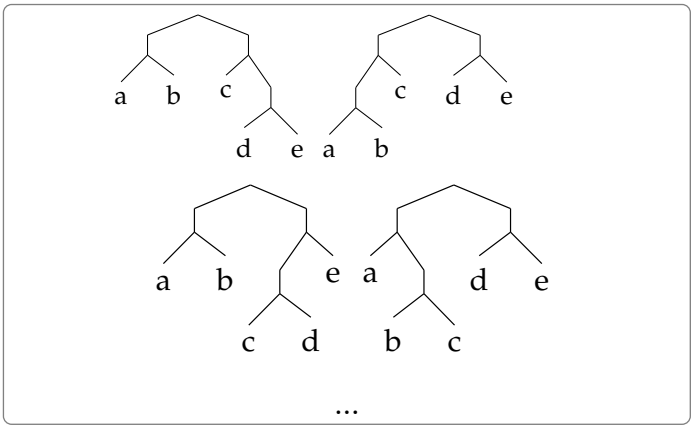
How many different binary trees can span a sentence of N words?



Combinatorial problems

A typical linguistic problem: parsing

How many different binary trees can span a sentence of N words?



words	trees
2	1
3	2
4	5
5	14
10	4862
20	1 767 263 190
...	...

Ambiguity and natural languages

with pretty pictures

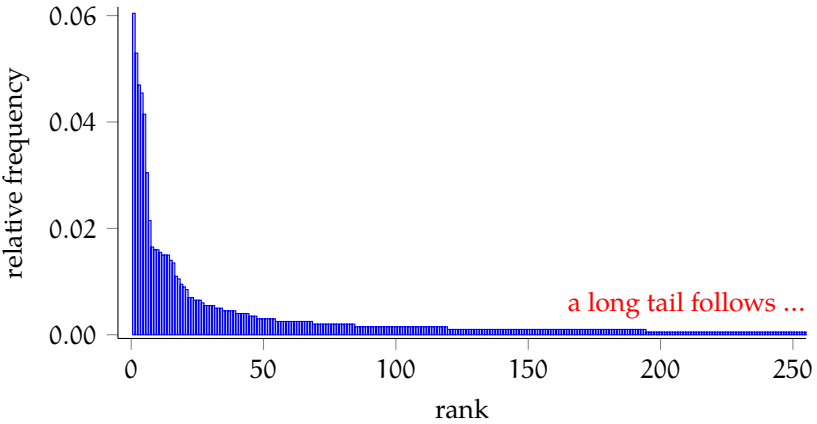
Cartoon Theories of Linguistics, SpecGram Vol CLIII, No 4, 2008. http://specgram.com/CLIII.4/school.gif

Statistical methods and data sparsity

- Statistical methods (machine learning) are the best way we know to deal with ambiguities
- Even for rule-based approaches, a statistical disambiguation component is often needed
- We need (annotated) data to learn, but ...

Languages are full of rare events

word frequencies in a small corpus

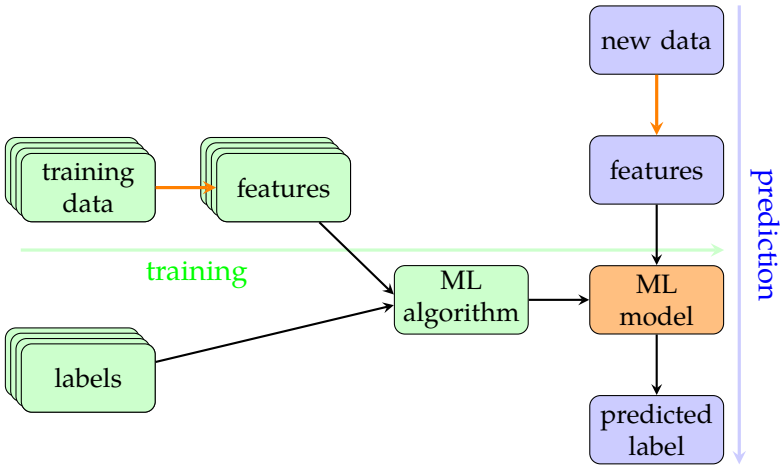


What is difficult in CL?

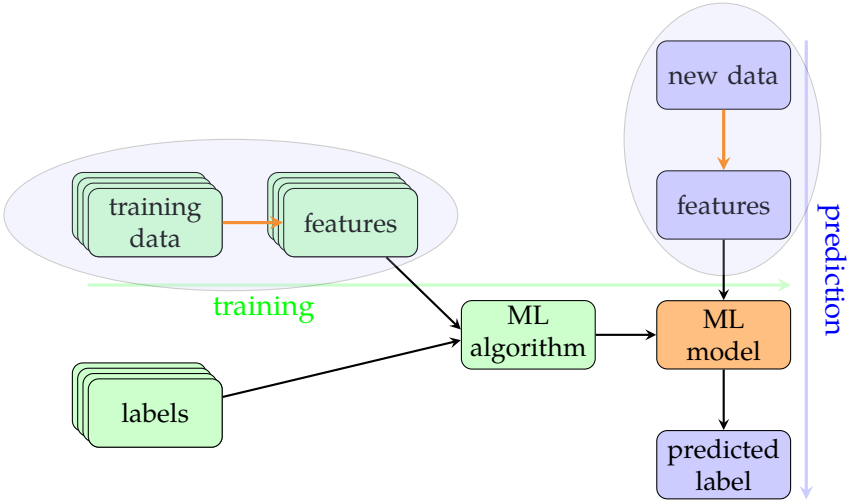
and how can machine learning help?

- Combinatorial problems - computational complexity
  - Often we resort to approximate methods: the answer to 'what is a good approximation?' comes from ML.
- Ambiguity
  - The answer to 'what is the best choice?' comes from ML.
- Data sparseness
  - Even here, ML can help.

Anatomy of a (supervised) data-driven solution



Anatomy of a (supervised) data-driven solution



What is in this course?

A bird's eye view

Introductory lectures on

- Quick recap of NLP 1
  - Math (linear algebra, calculus, probability/information theory)
  - Regression
  - Classification
  - *Generalization, bias, variance: Evaluation*
- Representation for language data
- Artificial neural networks
- Sequence models (HMMs, CRFs, RNNs)
- Unsupervised ANNs
- Language models
- Transformers and pretrained language models

Course overview

- Lectures
  - Wednesday 14:15-15:45 (VG 0.02)
  - Friday 14:15-15:45 (OSA 001)
- Labs
  - Mon 14:15-15:45 (VG 0.02) – starts from May 5
  - You will get hands on tutorials on a number tools
  - Also chance to ask questions on assignments
- Public course website: `https://snlp2-2025.github.io/`
- Moodle: `ttps://moodle.zdv.uni-tuebingen.de/course/view.php?id=774`
- GitHub: `https://github.com/snlp2-2025/snlp2`

Literature

- Mainly:
  Daniel Jurafsky and James H. Martin (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd. Online manuscript released January 12, 2025. URL: `https://web.stanford.edu/~jurafsky/slp3/`
- Reading recommendations from other sources will be provided for each course (all online, freely accessible)

Coursework and evaluation

- Reading material for most lectures
- Assignments: ungraded, but **required**
- Final (written) exam - (possibly) with a project component
- Attendance is not required, but you are unlikely to pass without regular attendance

Assignments

- 5 assignments, all will be released in two weeks through GitHub
- You are submit at least two different solutions for each
- We will likely reserve last two sessions of the lab for discussion of your solutions (with required attendance)
- The programming assignments can be done in pairs (recommended – knowing your classmates, and learning from them, is an important part of the university experience/education)
- This means **working together on the whole exercise**, not sharing parts of an assignment and working on them independently

Final remarks

- Please do not be shy, ask your questions during the lectures
- Please take the assignments seriously, learning programming requires practice
- Please fill in the 'beginning of semester survey' on Moodle
- Next:

Fri	Recap: math
Mon	Recap: regression

Final remarks

- Please do not be shy, ask your questions during the lectures
- Please take the assignments seriously, learning programming requires practice
- Please fill in the 'beginning of semester survey' on Moodle
- Next:

Fri	Recap: math
Mon	Recap: regression

Now, time for your questions

Acknowledgments, credits, references

- Chomsky, Noam (1968). "Quine's empirical assumptions". In: *Synthese* 19.1, pp. 53–68. DOI: 10.1007/BF00568049.
- Jurafsky, Daniel and James H. Martin (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd. Online manuscript released January 12, 2025. URL: https://web.stanford.edu/~jurafsky/slp3/.