Statistical Natural Language Processing

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

University of Tübingen
Seminar für Sprachwissenschaft

Summer Semester 2018

Why study (statistical) NLP

- (Most of) you are studying in a 'computational linguistics' program
- Many practical applications
- Investigating basic questions in linguistics and cognitive science (and more)

Application examples

For profit (engineering):

- Machine translation
- Question answering
- Information retrieval
- Dialog systems
- Summarization
- Text classification
- Text mining/analytics
- Sentiment analysis
- Speech recognition/synthesis
- Automatic grading
- Forensic linguistics

For fun (research):

- Modeling cognitive/social behavior
- Authorship attribution
- Investigating language change through time and space
- (Automatic) corpus annotation for linguistic research

Layers of linguistic analysis
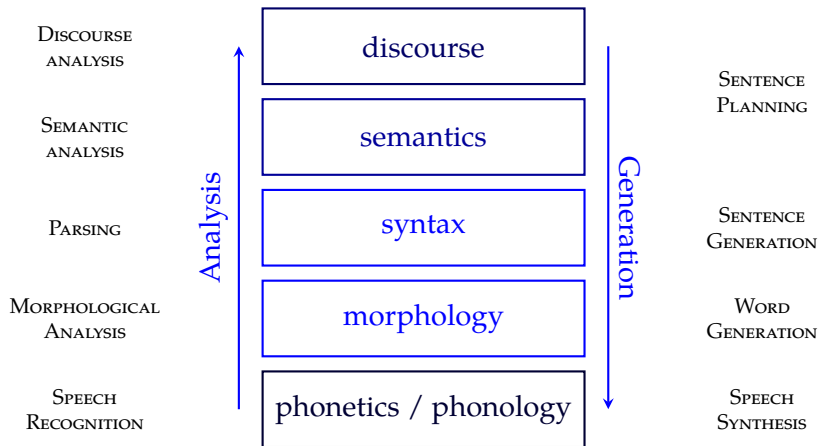
Annotation layers: example

From   the   AP   comes   this   story   :   →Tokens
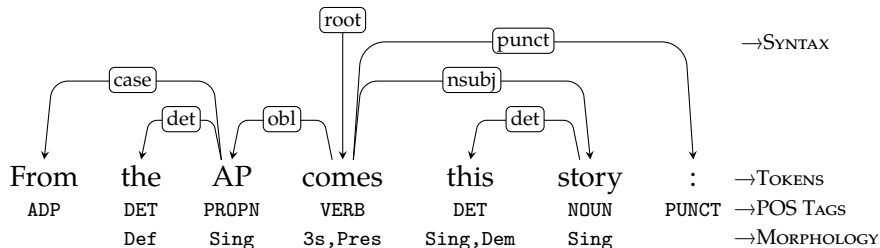
Annotation layers: example

From	the	AP	comes	this	story	:	→Tokens
ADP	DET	PROPN	VERB	DET	NOUN	PUNCT	→POS Tags
							→Morphology

Annotation layers: example

From	the	AP	comes	this	story	:	→Tokens
ADP	DET	PROPN	VERB	DET	NOUN	PUNCT	→POS Tags
	Def	Sing	3s,Pres	Sing,Dem	Sing		→Morphology

Annotation layers: example

Typical NLP pipeline

- Text processing / normalization
- Word/sentence tokenization
- POS tagging
- Morphological analysis
- Syntactic parsing
- Semantic parsing
- Named entity recognition
- Coreference resolution

Do we need a pipeline?

- Most "traditional" NLP architectures are based on a pipeline approach:
  - tasks are done individually, results are passed to upper level
- Joint learning (e.g., POS tagging and syntax) often improves the results
- End-to-end learning (without intermediate layers) is another (recent/trending) approach

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.* — **chomsky1968**

- Some linguistic traditions emphasize(d) 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 include some sort of statistical component

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 to calculate probability of sentence based on 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 to calculate probability of sentence based on the probabilities of words in it?

- Many similar questions we deal with have an exponential search space
- Naive approaches often are computationally intractable

NLP and ambiguity

fun with newspaper headlines

- FARMER BILL DIES IN HOUSE

NLP and ambiguity

fun with newspaper headlines

- FARMER BILL DIES IN HOUSE
- TEACHER STRIKES IDLE KIDS

NLP and ambiguity

fun with newspaper headlines

- FARMER BILL DIES IN HOUSE
- TEACHER STRIKES IDLE KIDS
- SQUAD HELPS DOG BITE VICTIM

NLP and ambiguity
fun with newspaper headlines

- FARMER BILL DIES IN HOUSE
- TEACHER STRIKES IDLE KIDS
- SQUAD HELPS DOG BITE VICTIM
- BAN ON NUDE DANCING ON GOVERNOR'S DESK

NLP and ambiguity
fun with newspaper headlines

- FARMER BILL DIES IN HOUSE
- TEACHER STRIKES IDLE KIDS
- SQUAD HELPS DOG BITE VICTIM
- BAN ON NUDE DANCING ON GOVERNOR'S DESK
- PROSTITUTES APPEAL TO POPE

NLP and ambiguity

fun with newspaper headlines

- FARMER BILL DIES IN HOUSE
- TEACHER STRIKES IDLE KIDS
- SQUAD HELPS DOG BITE VICTIM
- BAN ON NUDE DANCING ON GOVERNOR'S DESK
- PROSTITUTES APPEAL TO POPE
- KIDS MAKE NUTRITIOUS SNACKS

NLP and ambiguity
fun with newspaper headlines

- FARMER BILL DIES IN HOUSE
- TEACHER STRIKES IDLE KIDS
- SQUAD HELPS DOG BITE VICTIM
- BAN ON NUDE DANCING ON GOVERNOR'S DESK
- PROSTITUTES APPEAL TO POPE
- KIDS MAKE NUTRITIOUS SNACKS
- DRUNK GETS NINE MONTHS IN VIOLIN CASE

NLP and ambiguity

fun with newspaper headlines

- FARMER BILL DIES IN HOUSE
- TEACHER STRIKES IDLE KIDS
- SQUAD HELPS DOG BITE VICTIM
- BAN ON NUDE DANCING ON GOVERNOR'S DESK
- PROSTITUTES APPEAL TO POPE
- KIDS MAKE NUTRITIOUS SNACKS
- DRUNK GETS NINE MONTHS IN VIOLIN CASE
- MINERS REFUSE TO WORK AFTER DEATH

More ambiguities

we do not recognize many of them at first read

- Time flies like an arrow

- Outside of a dog, a book is
  a man's best friend

- One morning I shot an
  elephant in my pajamas

- Don't eat the pizza with
  knife and fork

More ambiguities

we do not recognize many of them at first read

- Time flies like an arrow;
  fruit flies like a banana.

- Outside of a dog, a book is
  a man's best friend

- One morning I shot an
  elephant in my pajamas

- Don't eat the pizza with
  knife and fork

More ambiguities

we do not recognize many of them at first read

- Time flies like an arrow;
  fruit flies like a banana.

- Outside of a dog, a book is
  a man's best friend; inside
  it's too hard to read.

- One morning I shot an
  elephant in my pajamas

- Don't eat the pizza with
  knife and fork

More ambiguities
we do not recognize many of them at first read

- Time flies like an arrow;
  fruit flies like a banana.

- Outside of a dog, a book is
  a man's best friend; inside
  it's too hard to read.

- One morning I shot an
  elephant in my pajamas.
  How he got in my pajamas,
  I don't know.

- Don't eat the pizza with
  knife and fork

More ambiguities

we do not recognize many of them at first read

- Time flies like an arrow;
  fruit flies like a banana.

- Outside of a dog, a book is
  a man's best friend; inside
  it's too hard to read.

- One morning I shot an
  elephant in my pajamas.
  How he got in my pajamas,
  I don't know.

- Don't eat the pizza with
  knife and fork ; the one
 with anchovies is better.

More ambiguities

we do not recognize many of them at first read

- Time flies like an arrow; fruit flies like a banana.

- Outside of a dog, a book is a man's best friend; inside it's too hard to read.

- One morning I shot an elephant in my pajamas. How he got in my pajamas, I don't know.

- Don't eat the pizza with knife and fork ; the one with anchovies is better.

- Hearing voices? Then you're not alone!

- No parking on both sides.

- They are canning peas.

- My job was keeping him alive.

- We watched another fly.

- Double job pay.

- He fed her cat food.

Even more ambiguities

with pretty pictures
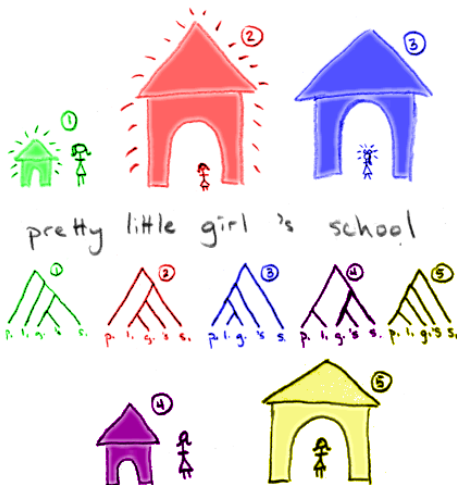
Statistical methods and data sparsity

- Statistical methods (machine learning) are the best way we know to deal with ambiguities
- Even for rule-based approaches, a statistical disambiguation component is necessary
- Machine learning methods require (annotated) data
- But …

Languages are full of rare events

word frequencies in a small corpus

a long tail follows …
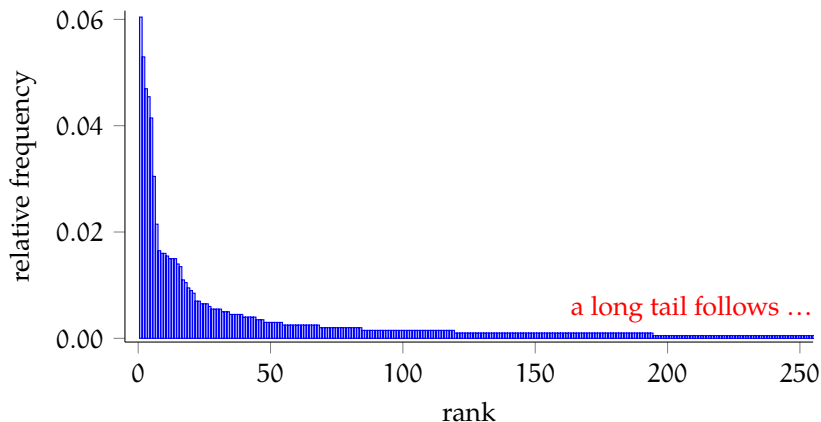
What is in this course

- Quick introduction / refreshers on important prerequisites
- The computational linguist's toolbox: basic methods and tools in NLP
- Some applications of NLP

What is in this course
Preliminaries

- Linear algebra, some concepts from calculus
- Probability theory
- Information theory
- Statistical inference
- Some topics from machine learning
  - Regression & classification
  - Sequence learning (HMMs)
  - Neural networks and deep learning
  - Unsupervised learning

What is in this course
NLP Tools and techniques

- Tokenization, normalization, segmentation
- N-gram language models
- Part of speech tagging
- Statistical parsing
- Distributed representations (of words, and other linguistic objects)

What is in this course

Applications

- Text classification
  - sentiment analysis
  - language detection
  - authorship attribution
  - …

What is in this course
Applications

- Text classification
  - sentiment analysis
  - language detection
  - authorship attribution
  - …

If time allows

- Statistical machine translation
- Named entitiy recognition
- Text summarization
- Dialog systems
- …

What is not in this course

- Cutting edge, latest methods & applications
- In-depth treatment of particular topics
- Introduction to terms / concepts from linguistics

Logistics

- Lectures: Mon/Fri 12:15 at Hörsaal 0.02
- Practical sessions: Wed 10:15 at Hörsaal 0.02
- Office hours: Wed 12:00-14:00 (room 1.09), or by appointment (email ccoltekin@sfs.uni-tuebingen.de)
- Course web page:
  http://sfs.uni-tuebingen.de/~ccoltekin/courses/snlp
- We will use GitHub classroom in this class (more on this soon)

Reading material

- **jurafsky2009**
  - Draft chapters of the third edition is available at
    `http://web.stanford.edu/~jurafsky/slp3/`

- **hastie2009**

Grading / evaluation

- Seven graded homework assignments (5 % each)
- Final exam (70 %)
- Attendance
  - 5 % (bonus) if you miss only one or two classes
  - you lose one bonus point for each additional class you miss
- Up to 5 % additional bonus points for Easter eggs:
  - first person finding intentional trivial mistakes in the course material gets 1 %

Assignments

- For distribution and submission of assignments, we will use GitHub Classroom

- The amount of git usage required is low, but learning/using git well is strongly recommended

- You are encouraged to pair up for the assignments, but you cannot pair with the same person twice

- Late assignments up to one week, will be graded up to half points indicated

- The solutions will be discussed in the tutorial session after one week from deadline

Assignment 0

- Your first assignment is already posted on the web page
- You need to follow the URL on the print version of the syllabus
- By completing assignment 0, you will
  - register for the course
  - have access to the non-public course material
  - exercise with how later assignments will work
  - provide some data for future exercises
- The repository created for assignment 0 is private, and can only be accessed you and the instructors

Practical sessions

- Tutor: Verena Blaschke
  ⟨verena.blaschke@student.uni-tuebingen.de⟩
- We will start with two sessions on Python
  tutorial/refresher
- You need to bring your own computer, make sure you have
  a working Python interpreter
- You are encouraged to ask questions about the exercises
  during practical sessions
- You are encouraged to ask questions about the assignments
- The solutions will be discussed during tutorial sessions

Further git/GitHub usage

- Once you complete Assignment 0, you will be a member of the 'organization' snlp2018
- You will get access to
  - private course material
  - assignment links
  - news and announcements

  through the repository at
  https://github.com/snlp2018/snlp2018
- Make sure to watch this repository
- You are also encouraged to use 'issues' in this repository as a place to discuss course topics, ask questions about the material and assignments

Next

Fri (this week) a hands-on introduction to Python

Mon Mathematical preliminaries (some linear algebra and bits from calculus)

Wed Python tutorial (continued)

References / additional reading material