Statistical Natural Language Processing

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

University of Tübingen
Seminar für Sprachwissenschaft

Summer Semester 2020

Why study (statistical) NLP

- (Most of) you are studying in a 'computational linguistics' program
- Many practical applications (NLP)
- Investigating basic scientific questions, primarily in linguistics and cognitive science (CL)

Application examples

Just a few examples

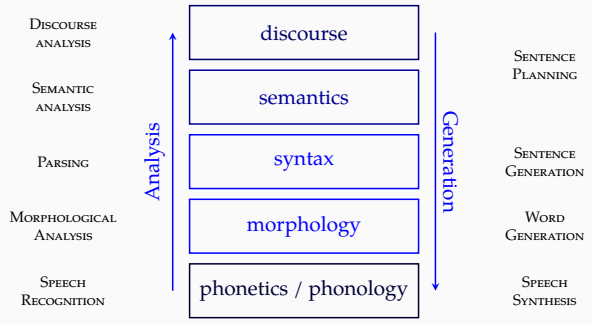
For profit (engineering):

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

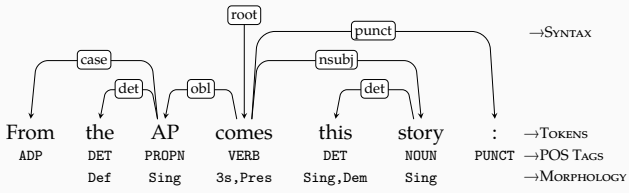
For fun (research):

- Modeling language processing learning
- Investigating language change through time and space
- Aiding language documentation through text processing
- Automatic corpus annotation for linguistic research
- Stylometry, author identification

Layers of linguistic analysis



Annotation layers: an example



Typical NLP pipeline

- Text processing / normalization
- Word/sentence tokenization, segmentation
- 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.* — *Chomsky (1968)*

- 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

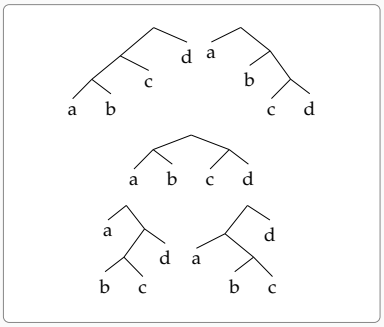
NLP and computational complexity

- How many possible parses a sentence may have?
- How many ways can you align two (parallel) sentences?
- How 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?



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

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
KIDS MAKE NUTRITIOUS SNACKS
DRUNK GETS NINE MONTHS IN VIOLIN CASE
MINERS REFUSE TO WORK AFTER DEATH
PROSTITUTES APPEAL TO POPE

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.

Even more ambiguities

with pretty pictures

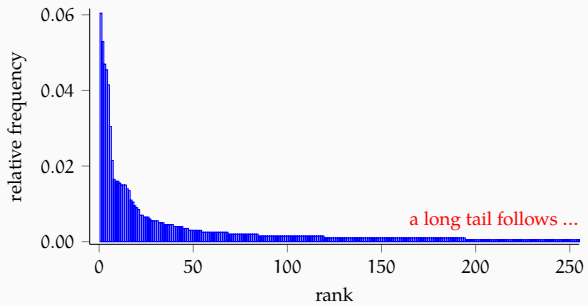
Cartoon Theories of Linguistics, SpecGram Vol CLIII, No 4, 2008. http://specgram.com/CLIII.4/school.gif

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.

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
  - Unsupervised learning
  - ... but what about 'deep learning'?
  - Short answer: we will cover the basics

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
  - ...

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 ~~in Hörsaal 0.02~~ online
- Practical sessions: Wed 10:15 ~~in Hörsaal 0.02~~ online
- Office hours: Mon 14:00-15:00 ~~(room 1.09)~~ by appointment (email ccoltekin@sfs.uni-tuebingen.de)
- Course web page: `https://snlp2020.github.io/`
- We will use GitHub classroom in this class (more on this soon)

Logistics

online classes

- Interaction is still important
  - Do not hesitate to ask questions during the online lectures
  - Asynchronous discussion via 'issues' at `https://github.com/snlp2020/snlp`
- This is new to all of us, we will learn how to handle it as we go
- Please provide feedback, suggestions – the instructors/tutors need it even more than before
- For of the lab sessions is particularly unclear

Reading material

- Daniel Jurafsky and James H. Martin (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. second. Pearson Prentice Hall. ISBN: 978-0-13-504196-3
  - Draft chapters of the third edition is available at `http://web.stanford.edu/~jurafsky/slp3/`
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second. Springer series in statistics. Springer-Verlag New York. ISBN: 9780387848587. URL: `http://web.stanford.edu/~hastie/ElemStatLearn/`
- Course notes for some lectures
- Other online references

Grading / evaluation

As a BA course (Proseminar)

- 7 graded assignments (6-best counts, 10 % each)
- Final exam (40 %)
- Quizzes with T/F or multiple choice questions (on Moodle)
  - Weekly, covering topics from the previous week
  - You have to get all questions correct
  - You have unlimited trials
  - If you complete all, you get 5 bonus points, each quiz missed reduces the bonus by one point
- Up to 5 % additional bonus points for Easter eggs:
  - first person finding mistakes in the course material gets 1 %
  - Easter eggs are intentionally placed, but you may also get bonus points for spotting unintentional mistakes

Grading / evaluation

For master's students

- You can take the class as a 'Proseminar' for 6ECTS, with the same requirements
- You can take the class as a 'Hauptseminar' (HS) for (only) 9ECTS with an additional project/paper related to the topics taught in the class
- If you choose the HS option, contact me with your project ideas as soon as you get some ideas

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 work on the assignments in pairs, but **you can work with the same person only once**
- 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
- We have a match-making system for working in random groups

Assignment 0

- Your first assignment is already posted on the web page
- By completing assignment 0, you will
  - register for the course
  - have access to the non-public course material
  - exercise with the way later assignments will work
  - provide some data for future exercises
- The repository created for assignment 0 is private, and can only be accessed by you and the instructors
- Please make sure that your assignment passes the tests (there are two 'pytest' tests in 'tests/' folder)

Practical sessions

- Tutor: Maximilian Gutsche
- Make sure you have a working Python interpreter
- Python 3 is strongly recommended
- You are encouraged to ask questions about the exercises during practical sessions
- The solutions will be discussed during tutorial sessions
- We need your opinions: how to hold lab sessions?

Further git/GitHub usage

- Once you complete Assignment 0, you will be a member of the 'organization' `snlp2020`
- You will get access to
  - private course material
  - assignment links
  - news and announcements

  through the repository at `https://github.com/snlp2020/snlp`
- Make sure you are watching 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 Mathematical preliminaries (some linear algebra and bits from calculus)

Mon Probability theory

References / additional reading material

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*. Springer. ISBN: 978-0387-31073-2.

Chomsky, Noam (1968). "Quine's empirical assumptions". In: *Synthese* 19.1, pp. 53–68. DOI: 10.1007/BF00568049.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second. Springer series in statistics. Springer-Verlag New York. ISBN: 9780387848587. URL: http://web.stanford.edu/~hastie/ElemStatLearn/.

Jurafsky, Daniel and James H. Martin (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. second. Pearson Prentice Hall. ISBN: 978-0-13-504196-3.

Manning, Christopher D. and Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. MIT Press. ISBN: 9780262133609.