Statistical Natural Language Processing

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

University of Tübingen
Seminar für Sprachwissenschaft

Summer Semester 2017

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

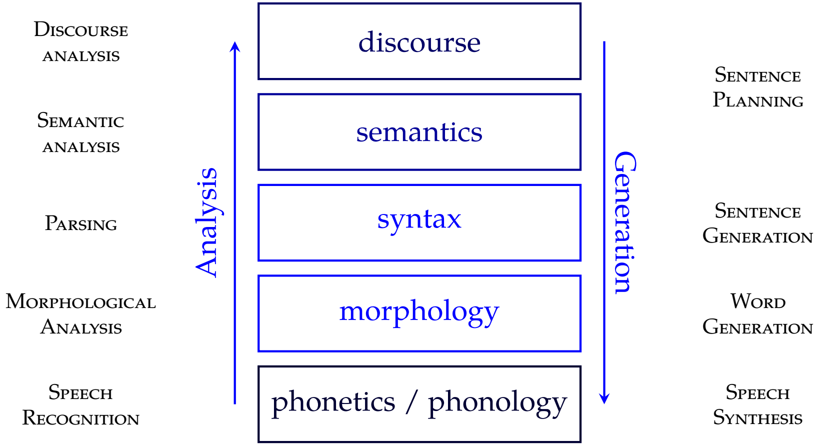
For profit (engineering):

- 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



Discourse analysis

Semantic analysis

Parsing

Morphological Analysis

Speech Recognition

Analysis

discourse

semantics

syntax

morphology

phonetics / phonology

Generation

Sentence Planning

Sentence Generation

Word Generation

Speech Synthesis

Annotation layers: example

From the AP comes this story : →TOKENS

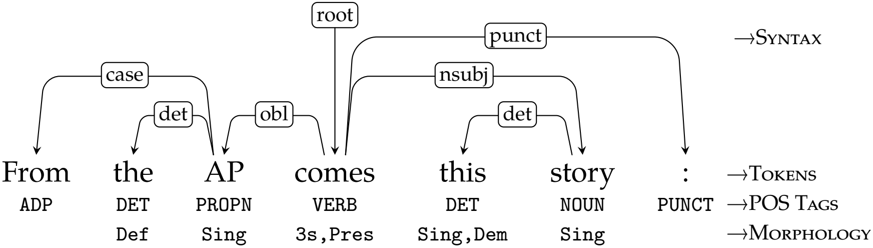
Annotation layers: example

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. — 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 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
- 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.

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
- 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.

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
- 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.

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
- 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.

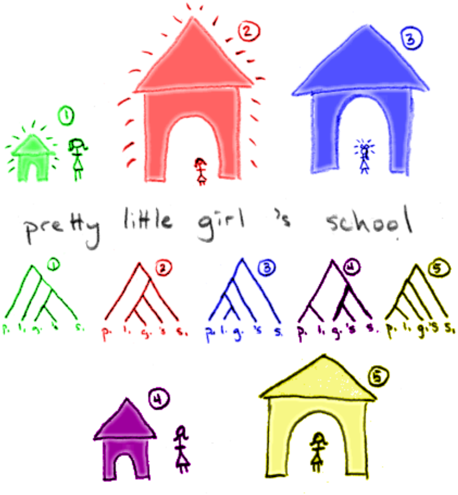
More ambiguities

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



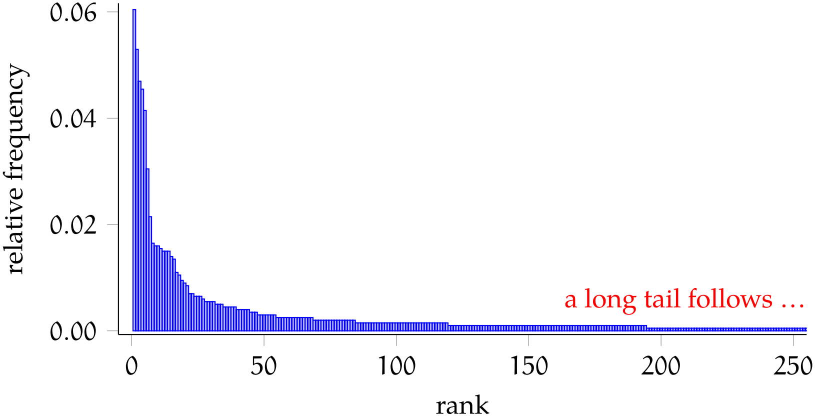
Cartoon Theories of Linguistics, SpecGram Vol CLIII, No 4, 2008. `http://specgram.com/CLIII.4/school.gif`

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 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
- Sequence alignment
- Distributed representations (of words, and other linguistic object)
- Text classification

What is in this course

Applications

- Statistical machine translation
- Sentiment analysis
- Topic models
- …

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/Wed/Fri 12:15 at Hörsaal 0.02
  Normally:

Mon/Wed	Formal lectures
Fri	Hands-on exercises

- Office hours: Wed 10:00-12:00 (room 1.09), or by appointment (email ccoltekin@sfs.uni-tuebingen.de)
- Course web page:
  `http://sfs.uni-tuebingen.de/~ccoltekin/courses/snlp`
- We also have a Moodle page (linked from the course web page)

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/`

Grading / evaluation

- Three graded homework assignments (10 % each)
- Final exam (70 %)
- Many non-graded (but not optional) exercises
- Attendance
  - 5 % (bonus) if you miss only one or two classes
  - you loose one 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 5 %

Practical sessions

- Tutor: Kuan Yu ⟨`kuan.yu@student.uni-tuebingen.de`⟩
- All programming exercises (graded or non-graded) should be done in Python
- The exercises are not graded, but they should not be considered optional

Next

Fri	(this week and next) a hands-on introduction to python
Mon	Mathematical preliminaries (some linear algebra and bits from calculus)
Wed	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.