Natural Language Processing

Introduction, course logistics. Part 2

Yulia Tsvetkov

yuliats@cs.washington.edu

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Syllabus https://courses.cs.washington.edu/courses/cse447/22sp/

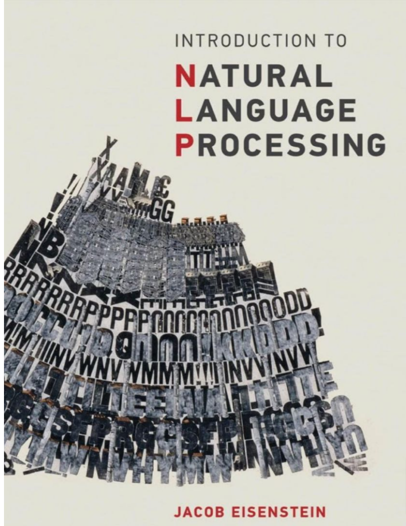
- Introduction
    - Overview of NLP as a field
- Modeling (ML fundamentals)
    - Text classification: linear models (perceptron, logistic regression), non-linear models (FF NNs, CNNs)
    - Language modeling: n-gram LMs, neural LMs, RNNs
    - Representation learning: word vectors, contextualized word embeddings, Transformers
- Linguistic structure and analysis (Algorithms, linguistic fundamentals)
    - Words, morphological analysis,
    - Sequences: part of speech tagging (POS), named entity recognition (NER)
    - Syntactic parsing (phrase structure, dependencies)
- Applications (Practical end-user solutions, research)
    - Sentiment analysis, toxicity detection
    - Machine translation, summarization
    - Computational social science
    - Interpretability
    - Fairness and bias

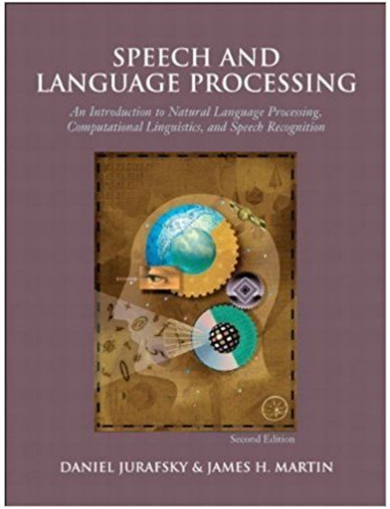
Course structure

please read the syllabus

https://courses.cs.washington.edu/courses/cse447/22sp/

Readings





- https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf
- https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf
- +additional readings posted weekly

Course website

- https://courses.cs.washington.edu/courses/cse447/22sp/
- Office hours, announcements, calendar, etc.

Deliverables & grading

- **Homework projects** - 90%
    - 3 programming assignments, 30% each
    - "Semi-autograded" – Most of the grades (~80%) come from replicating reference outputs in a given Jupyter notebook. You would usually know this part of your grades before submitting your assignments. The rest of the grades would involve things like write-ups, algorithm performance on hidden test sets, etc.
    - We'll discuss the setup in detail in the next lecture
- **Quizzes** - 10%
    - 8 simple quizzes on Wednesdays
    - 10 minutes in the beginning of the class
    - Starting from the 3rd week
    - 5 best quizzes, 2% each
- **Participation in course discussions** - 10% **bonus**
    - Respond to HW questions and discussions from your classmates
    - Contribute insightful discussions on Ed - 5% extra credit per 3 responses (10% max)

Homework assignments

- Project 1: Text classification
  - We will build a system for automatically classifying song lyrics comments by era. Specifically, we build machine learning text classifiers, including both generative and discriminative models, and explore techniques to improve the models.
- Project 2: Sequence labeling
  - We focus on sequence labeling with Hidden Markov Models and some simple deep learning based models. Our task is part-of-speech tagging on English and Norwegian from the Universal Dependencies dataset. We will cover the Viterbi algorithm which could require a little bit prior knowledge of dynamic programming.
- Project 3: Dependency parsing
  - We will implement a transition-based dependency parser. The algorithm would be new and specific to the dependency parsing problem, but the underlying building blocks of the method are still some neural network modules covered in P1 and P2.

Homework submission

- Submit via Gitlab
  - We will pull your code for submission (with an assignment tag) and check the commit time.
  - A detailed grading rubric would be specified in the main Jupyter notebook of each assignment.
- Late policy
  - Each student will be granted **5 late days** to use over the duration of the quarter.
  - You can use a **maximum of 3 late days on any one project**.
  - Weekends and holidays are also counted as late days.
  - Late submissions are automatically considered as using late days.
  - Using late days will not affect your grade.
  - However, projects submitted late after all late days have been used will receive no credit. Be careful!

Communications with instructors

- You should be able to see yourselves be added to the Ed discussion board of CSE 447 / CSE M 547 22 sp. **Please contact the staff if you are not.**
- **Discussion Board (EdSTEM)** will be used to answer questions related to lectures and assignments
  - We really encourage you to ask/discuss higher level questions on the discussion board.
  - We encourage that generic questions should be posted as “Public” so that other classmates would also got benefited from it.
  - Please do not post detail about your solutions (detail ideas, codes, etc.) on public threads. Private discussion should be used for these posts.
- For grading issues, please email the instructor team directly.

Class participation

- Lectures and homework assignments complement each other
- Lecture materials are broader
- Homework assignments will go deeper into three important topics
- Try to attend the lectures
- Quizzes are designed to encourage you to do so
- But if you miss a lecture – you can watch a recording, or read assigned book chapters
- Participate in class discussions, 10% bonus is an incentive
  - But don’t just provide code solutions to questions on homework projects– those are for individual work!
  - Provide insights, theoretical background, references to readings
- Your questions are always welcome!

Office hours

- Yulia – Thu 3:30 - 4:30pm @ zoom https://washington.zoom.us/j/92241001365 (and by appointment)
- Thai - Tues 2:00 - 3:00pm @ Zoom (Zoom ID on Canvas)
- Han - Mon 11am-12pm @ zoom (https://washington.zoom.us/my/xhan77)
- Kaiser - Mon 12:20pm - 1:20pm @ zoom (Zoom ID on Canvas)
- Ivy - Wed 4:30-5:30pm (https://washington.zoom.us/my/ivyguo)
- Leroy - Fri 11:30am - 12:30pm (https://washington.zoom.us/my/lrywng)

Quizzes

- 8 quizzes, students can drop 3
- Each quiz has 5 simple multiple-choice questions, autograded
- Quizzes are on Canvas, open during the lecture time
- Quiz time - 10 minutes in the beginning of the class
- Starting from the 3rd week
- Grading on 5 best quizzes, 2% each

More course logistics

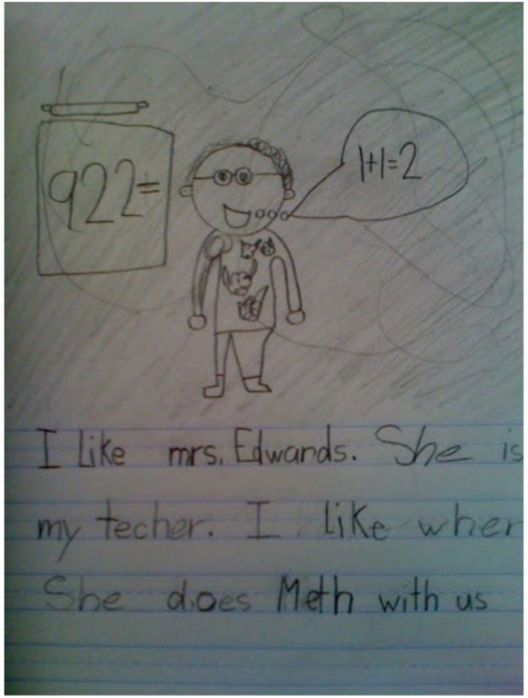
We care that you learn!

Your questions are always welcome.

Questions?

https://courses.cs.washington.edu/courses/cse447/22sp/

What does it mean to “know” a language?

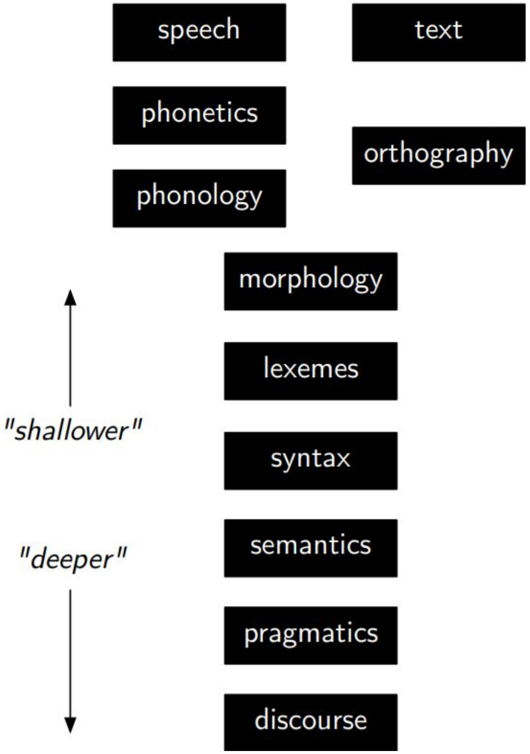


Hi, how can I help?

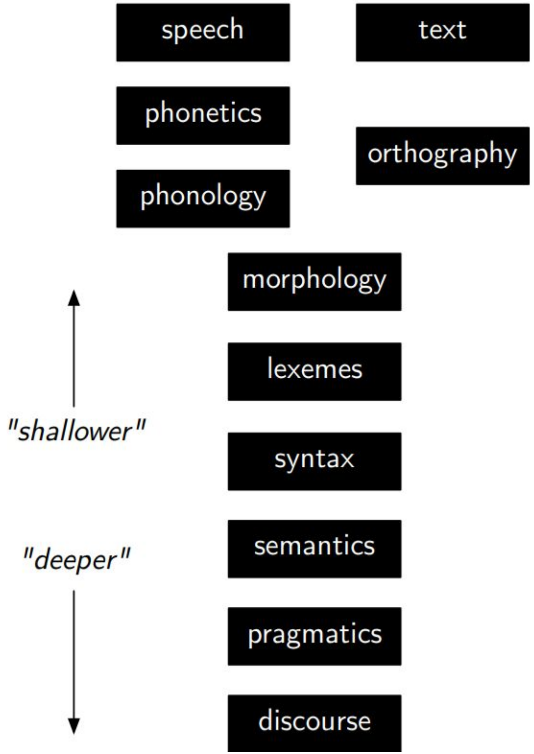
What does an NLP system need to 'know'?

- Language consists of many levels of structure
- Humans fluently integrate all of these in producing/understanding language
- Ideally, so would a computer!

Levels of linguistic knowledge



Speech, phonetics, phonology



Orthography

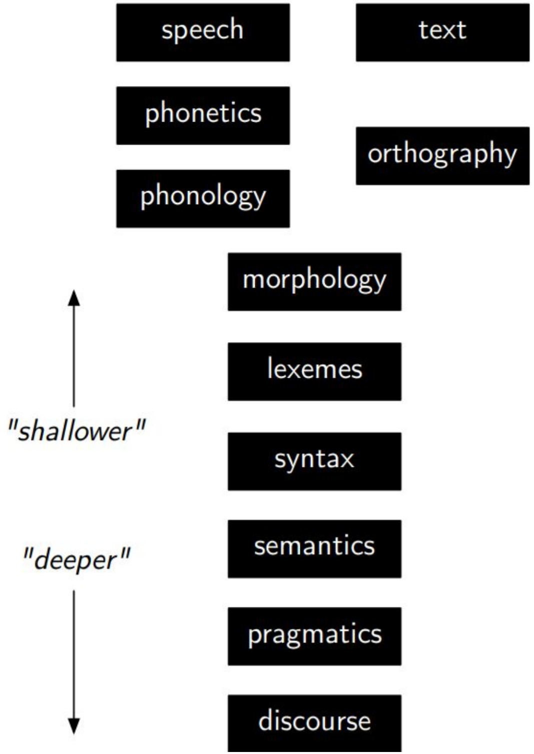
هذه جملة بسيطة

đây là một câu đơn giản

यह एक साधारण वाक्य है

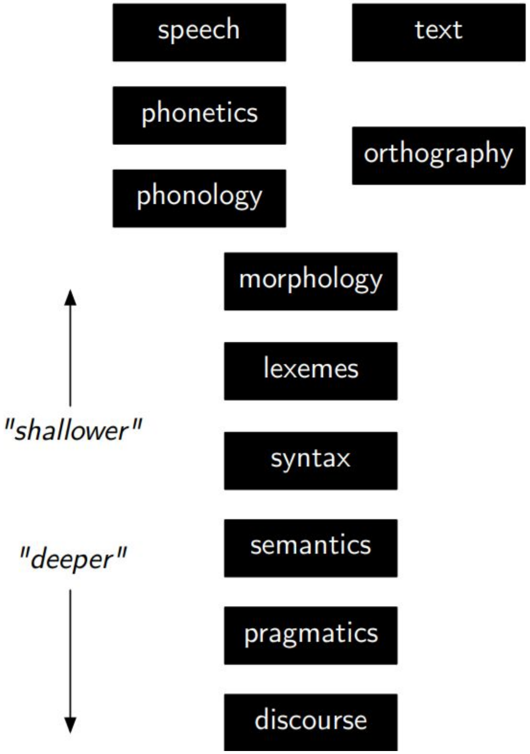
This is a simple sentence .

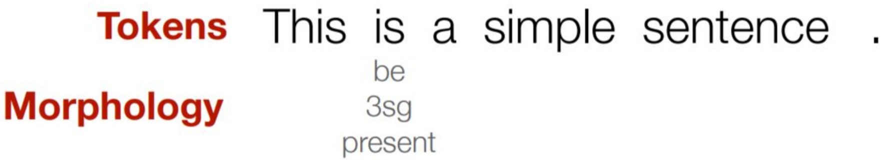
/ ðɪs ɪz ə ˈsɪmpl ˈsɛntəns /.



Words, morphology

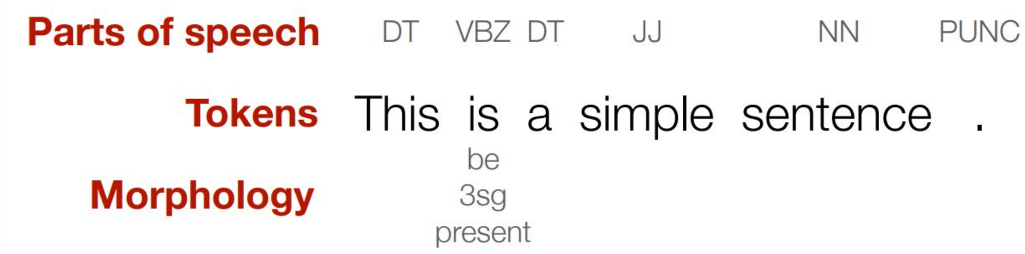
- Morphological analysis
- Tokenization
- Lemmatization

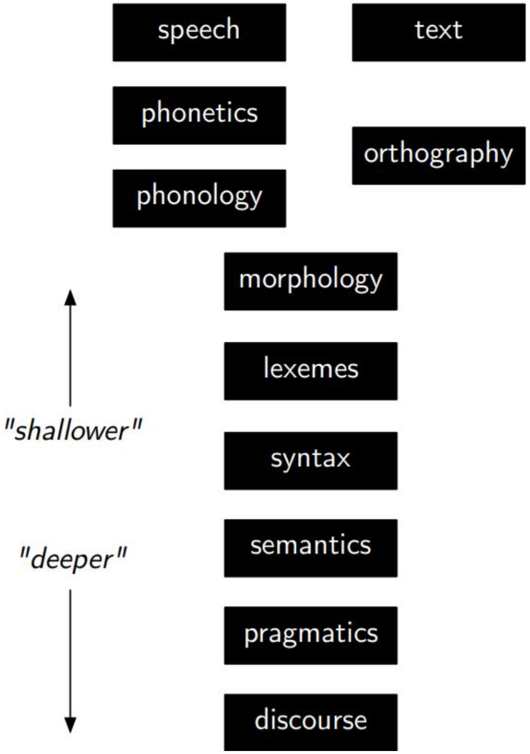




Syntax

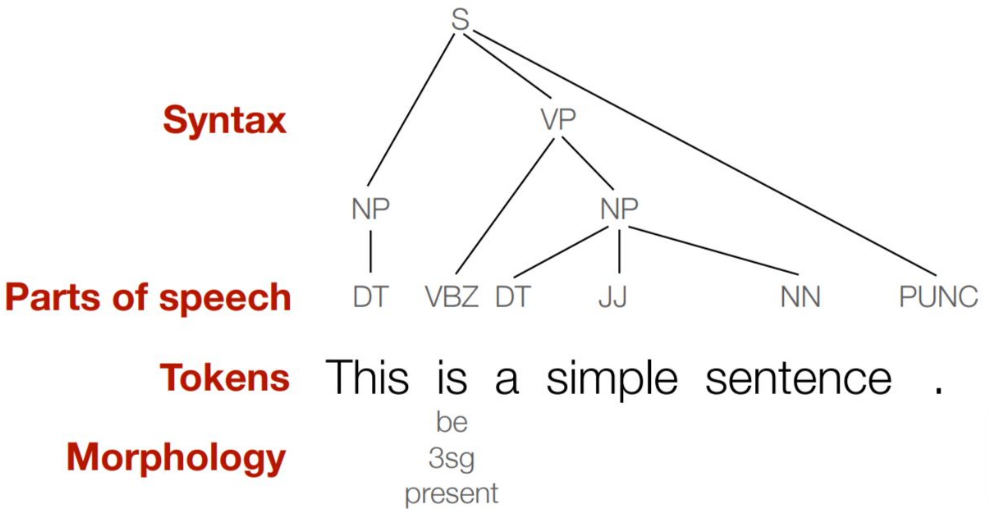
- Part-of-speech tagging

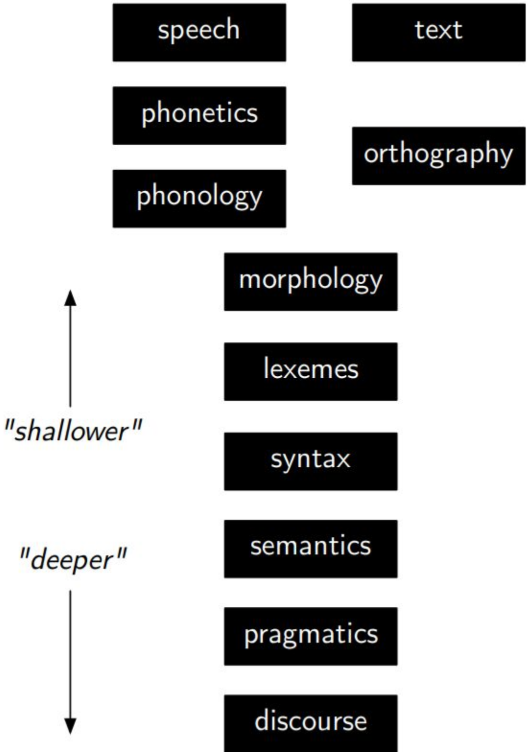




Syntax

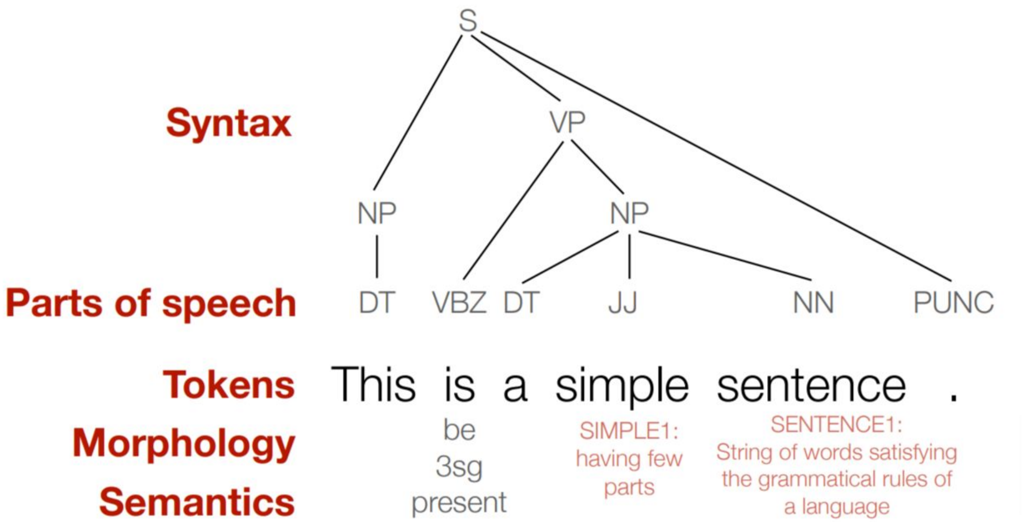
- Part-of-speech tagging
- Syntactic parsing

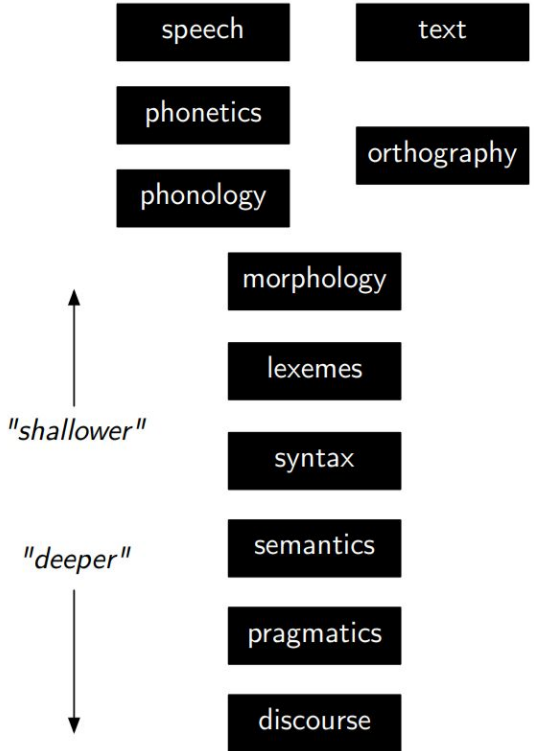




Semantics

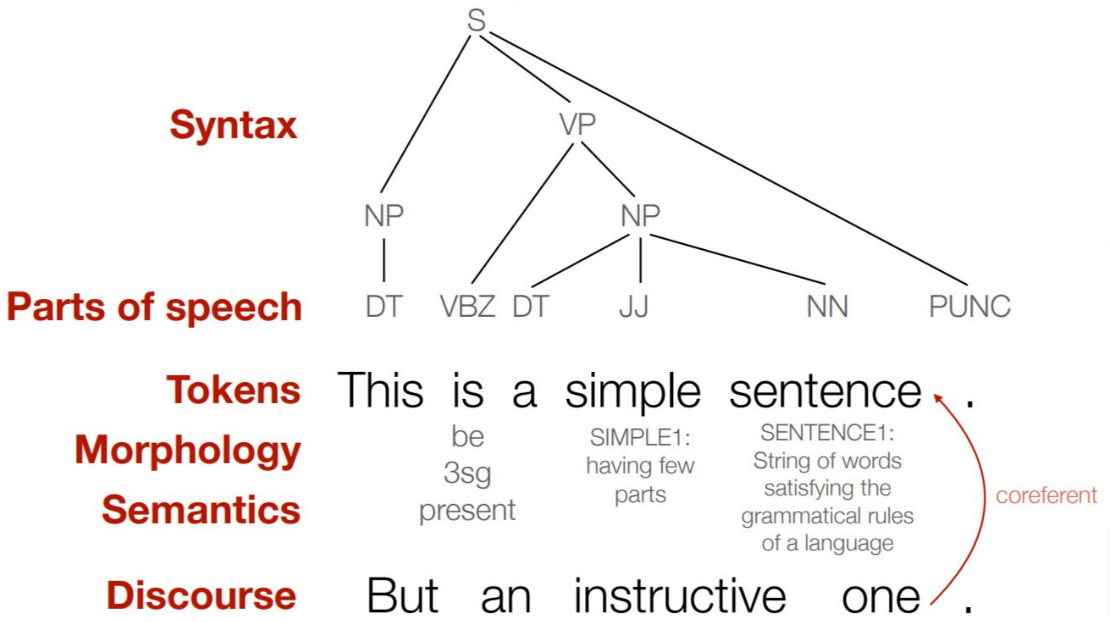
- Named entity recognition
- Word sense disambiguation
- Semantic role labelling

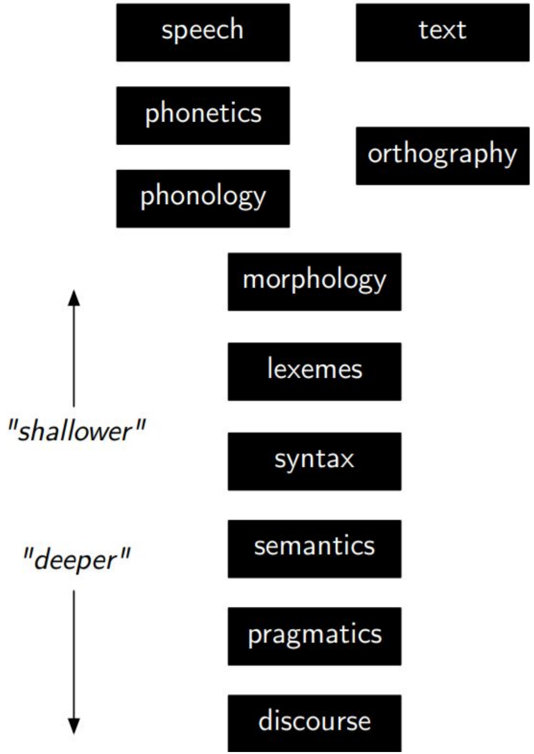




Discourse

- Reference resolution
- Discourse parsing

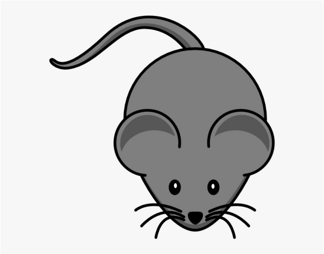


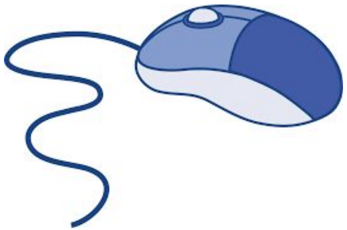


Why is language interpretation hard?

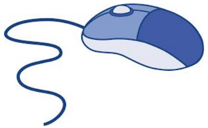
1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. Unknown representation $\mathcal{R}$

Ambiguity: word sense disambiguation





Ambiguity

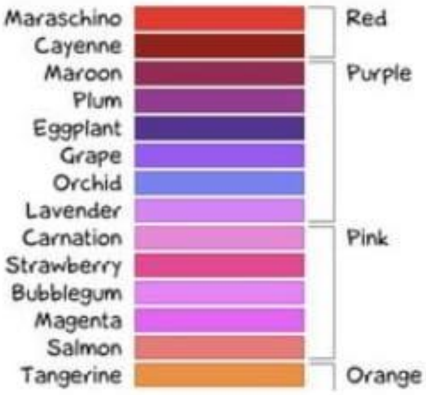



- Ambiguity at multiple levels:
  - Word senses: **bank** (finance or river?)
  - Part of speech: **chair** (noun or verb?)
  - Syntactic structure: **I can see a man with a telescope**
  - Multiple: **I saw her duck**



Semantic analysis

- Every language sees the world in a different way
  - For example, it could depend on cultural or historical conditions



  - Russian has very few words for colors, Japanese has hundreds
  - Multiword expressions, e.g. happy as a clam, it's raining cats and dogs or wake up and metaphors, e.g. love is a journey are very different across languages

Why is language interpretation hard?

1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. Unknown representation $\mathcal{R}$

Scale

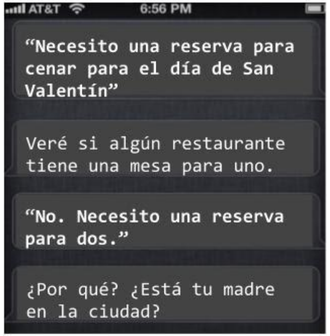
- ~7K languages
- Thousands of language varieties

Englishes

Africa is a continent with a very high linguistic diversity: there are an estimated 1.5-2K African languages from 6 language families. 1.33 billion people

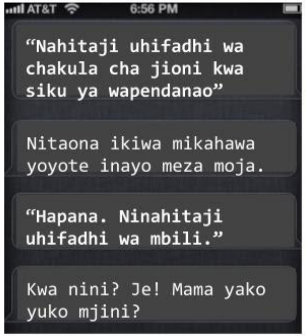
NLP beyond English

- ~7,000 languages
- thousands of language varieties



Spanish
534 million speakers

Hindi
615 million speakers



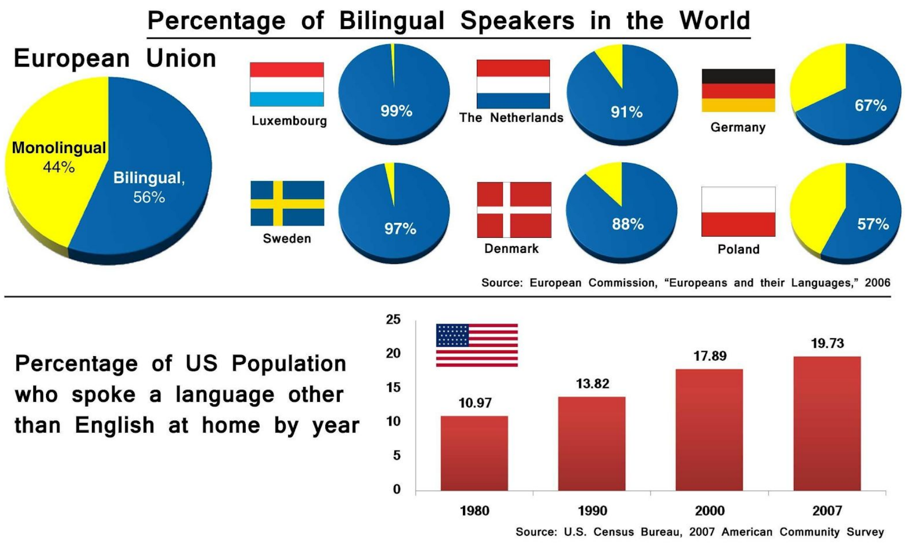
Swahili
100 million speakers

American English

Scottish English

Hinglish

Most of the world today is multilingual



Source: US Census Bureau

Source: Ethnologue

Tokenization

这是一个简单的句子

WORDS This is a simple sentence

זה משפט פשוט

Tokenization + disambiguation

in tea	בתה
her daughter	

- most of the vowels unspecified

in tea	בתה
in the tea	בהתה
that in tea	שבתה
that in the tea	שבהתה
and that in the tea	ושבהתה

ושבתה

and her saturday	ו+שבת+ה
and that in tea	ו+ש+ב+תה
and that her daughter	ו+ש+בת+ה

- most of the vowels unspecified
- particles, prepositions, the definite article, conjunctions attach to the words which follow them
- tokenization is highly ambiguous

Tokenization + morphological analysis

- Quechua

Much’ananayakapushasqakupuniñataqsunamá

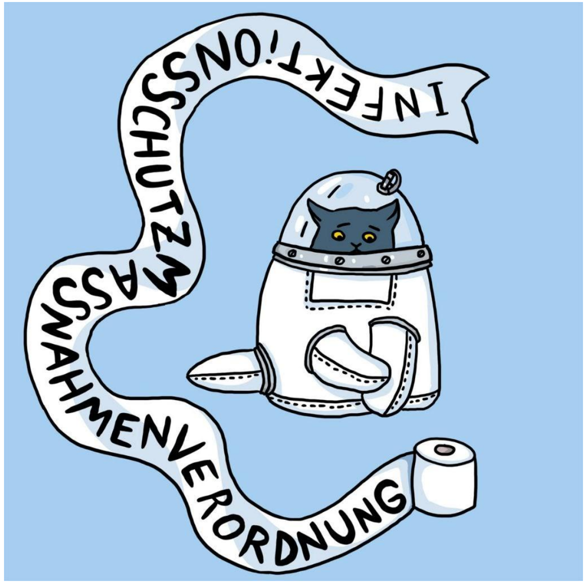
Much’a -na -naya -ka -pu -sha -sqa -ku -puni -ña -taq -suna -má

“So they really always have been kissing each other then”

Much’a	to kiss
-na	expresses obligation, lost in translation
-naya	expresses desire
-ka	diminutive
-pu	reflexive (kiss *eachother*)
-sha	progressive (kiss*ing*)
-sqa	declaring something the speaker has not personally witnessed
-ku	3rd person plural (they kiss)
-puni	definitive (really*)
-ña	always
-taq	statement of contrast (...then)
-suna	expressing uncertainty (So...)
-má	expressing that the speaker is surprised

Tokenization + morphological analysis

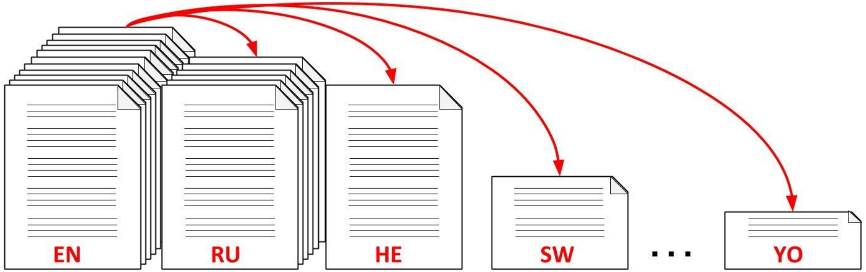
- German

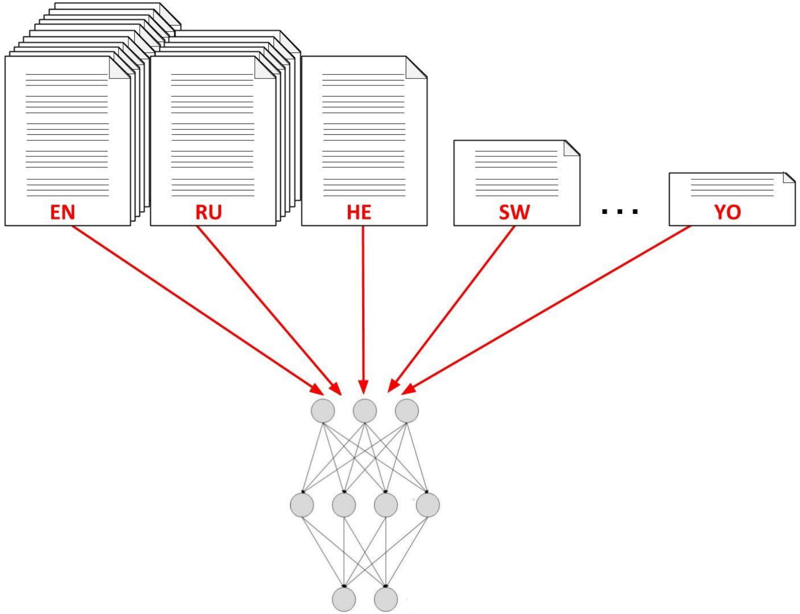


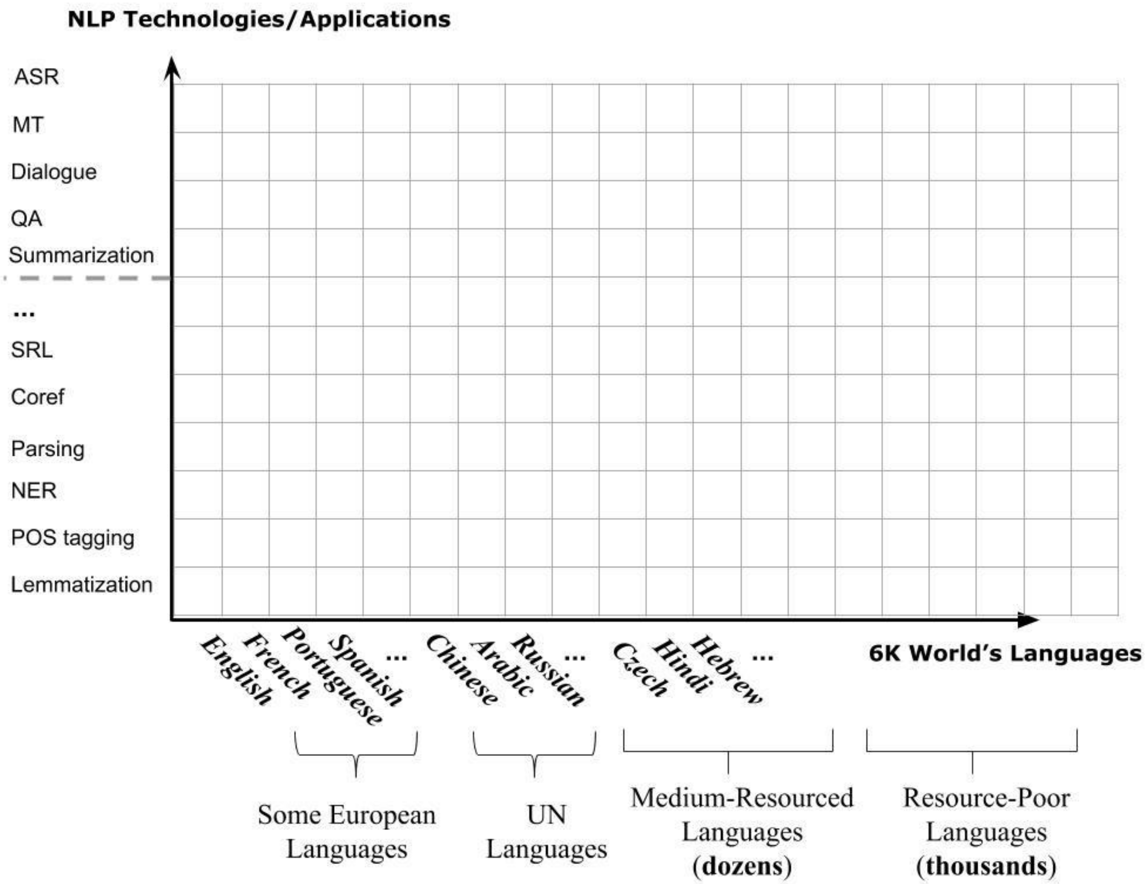
Infektionsschutzmaßnahmenverordnung

Multilingual NLP

- Levels of linguistic structure
- Categorization of languages and processing of linguistic structures across languages
- Multilingual modeling






NLP Technologies/Applications
ASR
MT
Dialogue
QA
Summarization
...
SRL
Coref
Parsing
NER
POS tagging
Lemmatization
English
French
Portuguese
Spanish
...
Chinese
Arabic
Russian
...
Czech
Hindi
Hebrew
...
6K World's Languages
Some European Languages
UN Languages
Medium-Resourced Languages (dozens)
Resource-Poor Languages (thousands)

Why is language interpretation hard?

1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. Unknown representation $\mathcal{R}$

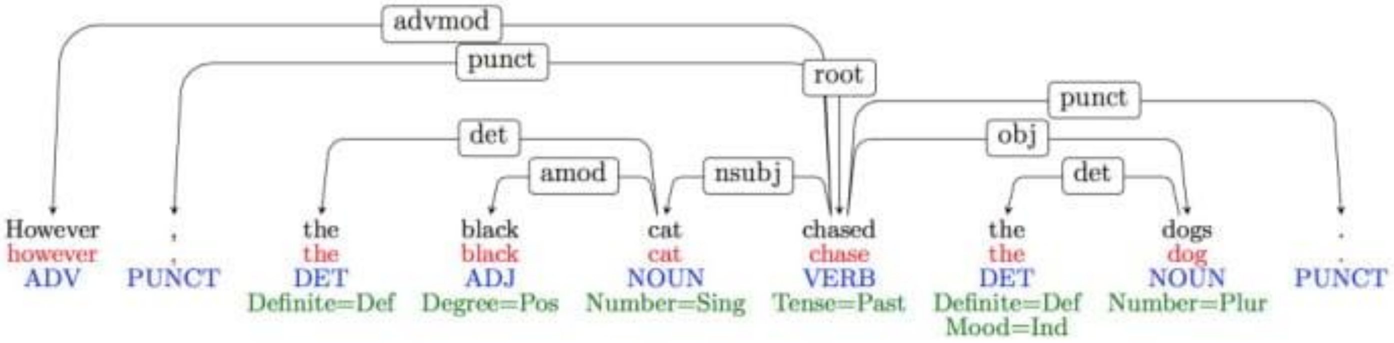
Linguistic variation

- Non-standard language, emojis, hashtags, names

chowdownwithchan #crab and #pork #xiaolongbao at @dintaifungusa... where else? 😂 🤷‍♀️ Note the cute little crab indicator in the 2nd pic 🦀 💕

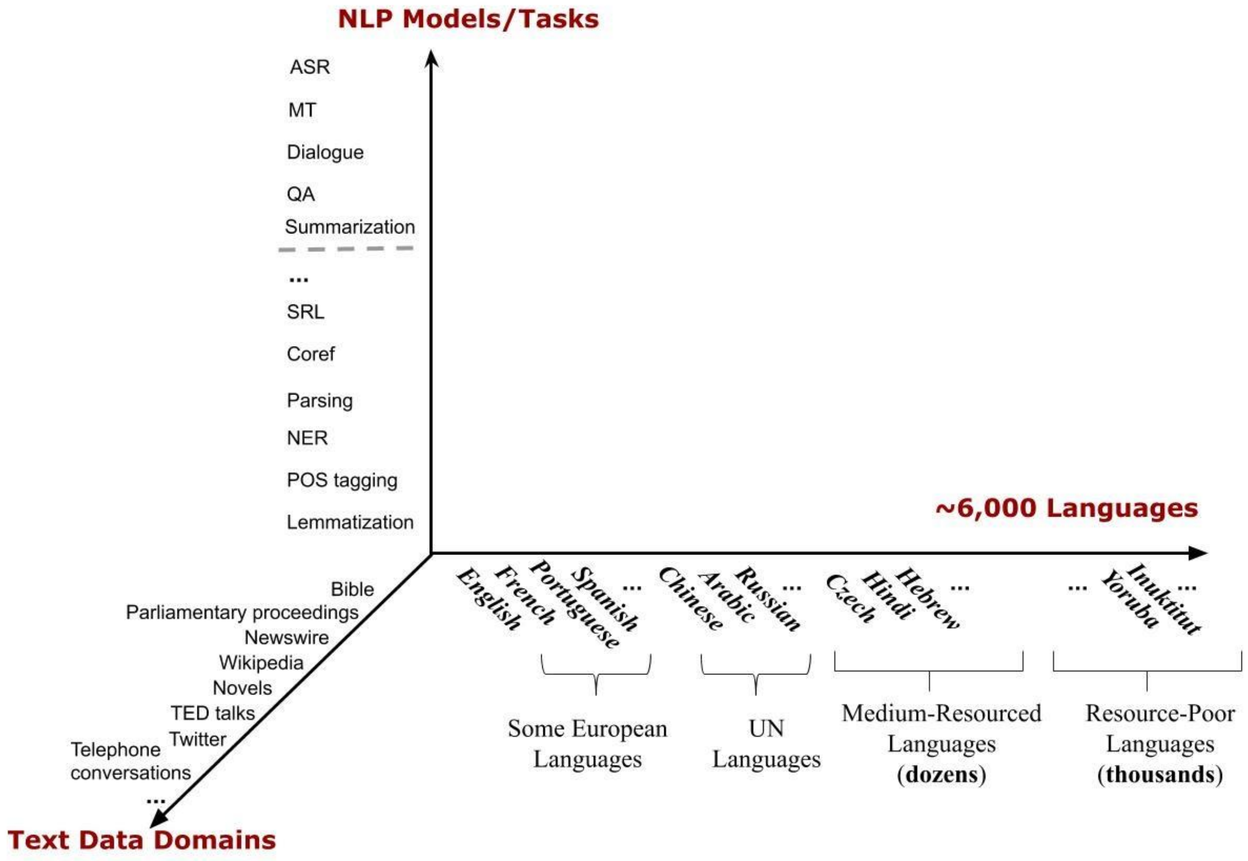
Variation

- Suppose we train a part of speech tagger or a parser on the Wall Street Journal



- What will happen if we try to use this tagger/parser for social media??

@_rkpntrnte hindi ko alam babe eh, absent ako kanina I'm sick rn hahaha 😌🙌


NLP Models/Tasks
ASR
MT
Dialogue
QA
Summarization
...
SRL
Coref
Parsing
NER
POS tagging
Lemmatization
~6,000 Languages
English
French
Portuguese
Spanish
...
Chinese
Arabic
Russian
...
Czech
Hindi
Hebrew
...
...
Yoruba
Inuktitut
...
Some European Languages
UN Languages
Medium-Resourced Languages (dozens)
Resource-Poor Languages (thousands)
Bible
Parliamentary proceedings
Newswire
Wikipedia
Novels
TED talks
Twitter
Telephone conversations
...
Text Data Domains

Why is language interpretation hard?

1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. Unknown representation $\mathcal{R}$

Sparsity

Sparse data due to Zipf's Law

- To illustrate, let's look at the frequencies of different words in a large text corpus
- Assume "word" is a string of letters separated by spaces

Word Counts

Most frequent words in the English Europarl corpus (out of 24m word tokens)

any word		nouns	
Frequency	Token	Frequency	Token
1,698,599	the	124,598	European
849,256	of	104,325	Mr
793,731	to	92,195	Commission
640,257	and	66,781	President
508,560	in	62,867	Parliament
407,638	that	57,804	Union
400,467	is	53,683	report
394,778	a	53,547	Council
263,040	I	45,842	States

Word Counts

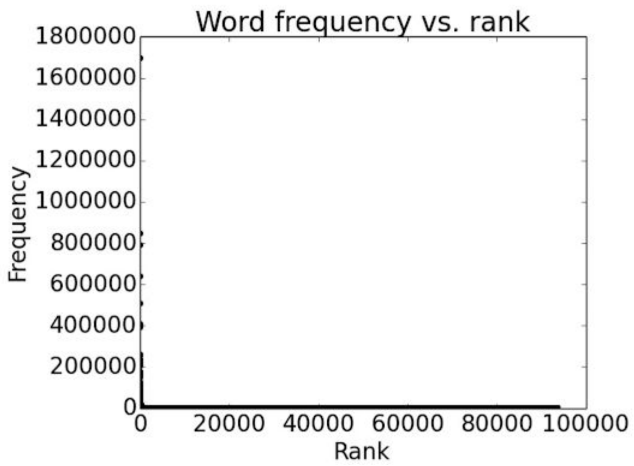
But also, out of 93,638 distinct words (word types), 36,231 occur only once.

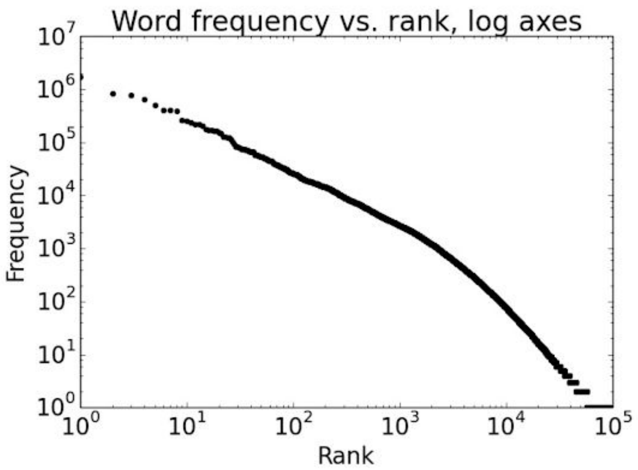
Examples:

- cornflakes, mathematicians, fuzziness, jumbling
- pseudo-rapporteur, lobby-ridden, perfunctorily,
- Lycketoft, UNCITRAL, H-0695
- policyfor, Commissioneris, 145.95, 27a

Plotting word frequencies

Order words by frequency. What is the frequency of nth ranked word?

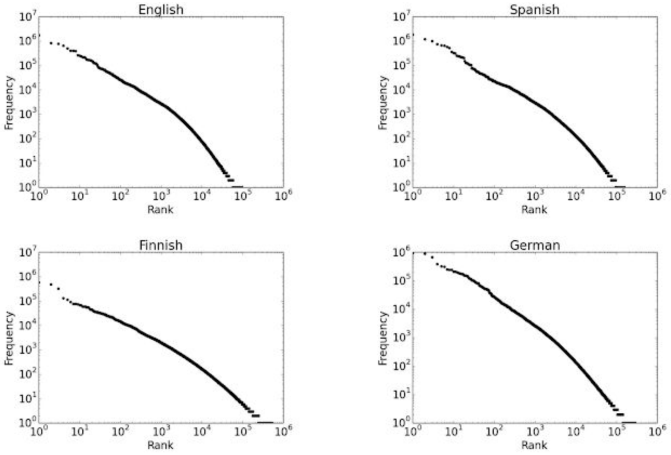




Zipf’s Law

Implications

- Regardless of how large our corpus is, there will be a lot of infrequent (and zero-frequency!) words
- This means we need to find clever ways to estimate probabilities for things we have rarely or never seen



Why is language interpretation hard?

1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. Unknown representation $\mathcal{R}$

Expressivity

Not only can one form have different meanings (ambiguity) but the same meaning can be expressed with different forms:

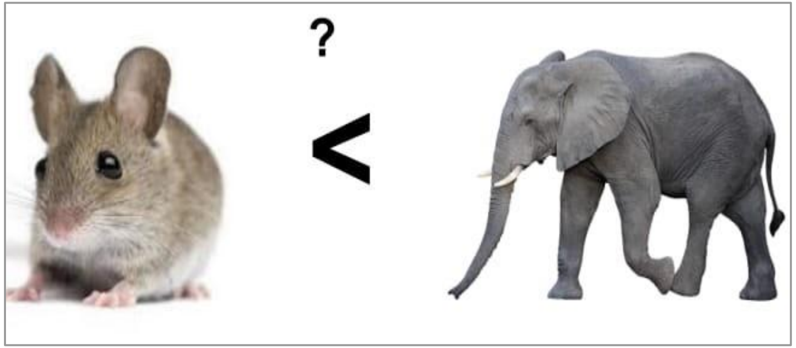
She gave the book to Tom	vs.	She gave Tom the book
Some kids popped by	vs.	A few children visited
Is that window still open?	vs.	Please close the window

Why is language interpretation hard?

1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. Unknown representation $\mathcal{R}$

Unmodeled variables

"Drink this milk"



World knowledge

- I dropped the glass on the floor and it broke
- I dropped the hammer on the glass and it broke

Why is language interpretation hard?

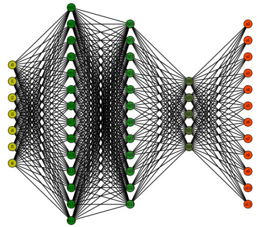
1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. Unknown representation $\mathcal{R}$

Unknown representation

- Very difficult to capture what is $\mathcal{R}$, since we don't even know how to represent the knowledge a human has/needs:
  - What is the "meaning" of a word or sentence?
  - How to model context?
  - Other general knowledge?

Dealing with ambiguity

- How can we model ambiguity and choose the correct analysis in context?
  - non-probabilistic methods (FSMs for morphology, CKY parsers for syntax) return *all possible analyses*.
  - probabilistic models (HMMs for part-of-speech tagging, PCFGs for syntax) and algorithms (Viterbi, probabilistic CKY) return *the best possible analysis,* i.e., the most probable one according to the model
  - Neural networks, pretrained language models now provide end-to-end solutions



- But the “best” analysis is only good if our probabilities are accurate. Where do they come from?

Corpora

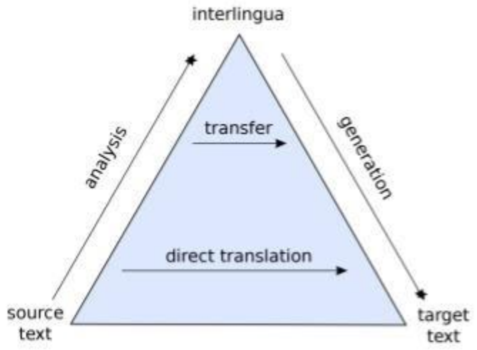
- A corpus is a collection of text
  - Often annotated in some way
  - Sometimes just lots of text
- Examples
  - Penn Treebank: 1M words of parsed WSJ
  - Canadian Hansards: 10M+ words of aligned French / English sentences
  - Yelp reviews
  - The Web: billions of words of who knows what

Desiderata for NLP models

- Sensitivity to a wide range of phenomena and constraints in human language
- Generality across languages, modalities, genres, styles
- Strong formal guarantees (e.g., convergence, statistical efficiency, consistency)
- High accuracy when judged against expert annotations or test data
- Ethical

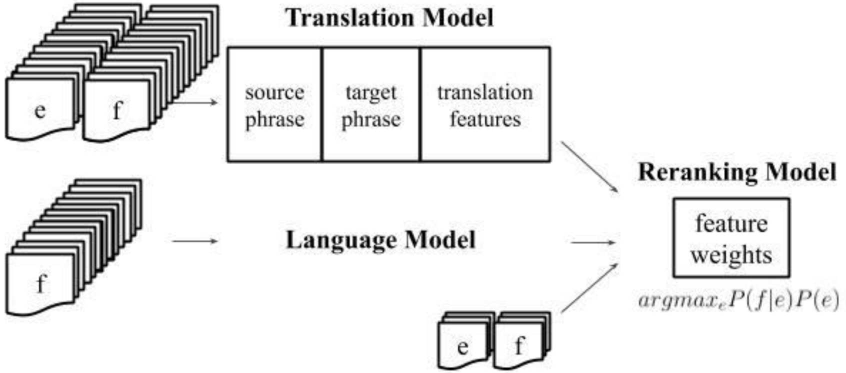
Symbolic and Probabilistic NLP

Logic-based/Rule-based NLP



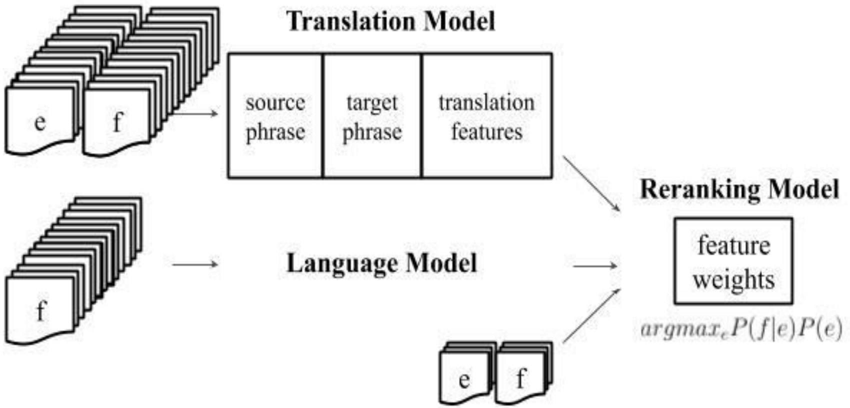
~ 90s

Statistical NLP



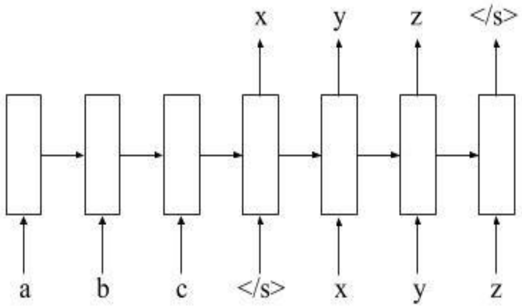
Probabilistic and Connectionist NLP

Engineered Features/Representations

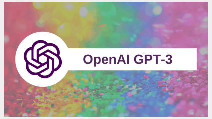


~mid 2010s

Learned Features/Representations



GPT-3


OpenAI GPT-3


weibo
twitter
facebook
YouTube
Google
LinkedIn
Google+
bing
AFP
AP
YAHOO!
CNN
myspace
phpBB
BBC NEWS
flickr
digg
Bloomberg
The New York Times
AI

Chat bots

REPORT

OPENAI'S LATEST BREAKTHROUGH IS ASTONISHINGLY POWERFUL, BUT STILL FIGHTING ITS FLAWS

The ultimate autocomplete

By James Vincent | Jul 30, 2020, 10:01am EDT

```
Q: How many eyes does a giraffe have?
A: A giraffe has two eyes.

Q: How many eyes does my foot have?
A: Your foot has two eyes.

Q: Which is heavier, a mouse or a toaster?
A: A toaster is heavier than a mouse.

Q: Which is heavier, a toaster or a pencil?
A: A pencil is heavier than a toaster.
```

OpenAI's new language generator GPT-3 is shockingly good—and completely mindless

The AI is the largest language model ever created and can generate amazing human-like text on demand but won't bring us closer to true intelligence.

by **Will Douglas Heaven** July 20, 2020

https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html

Where are we now? GPT-3 Turing Test

This gives us a hint for how to stump the AI more consistently. We need to ask questions that no normal human would ever talk about.

```
Q: How many eyes does a giraffe have?
A: A giraffe has two eyes.

Q: How many eyes does my foot have?
A: Your foot has two eyes.

Q: How many eyes does a spider have?
A: A spider has eight eyes.

Q: How many eyes does the sun have?
A: The sun has one eye.

Q: How many eyes does a blade of grass have?
A: A blade of grass has one eye.
```

https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html

GPT-3 Turing Test

Now we're getting into surreal territory. GPT-3 knows how to have a normal conversation. It doesn't quite know how to say "Wait a moment... your question is nonsense." It also doesn't know how to say "I don't know."

```
Q: How do you sporgle a morgle?
A: You sporgle a morgle by using a sporgle.

Q: How many bonks are in a quoit?
A: There are three bonks in a quoit.

Q: How many rainbows does it take to jump from Hawaii to seventeen?
A: It takes two rainbows to jump from Hawaii to seventeen.

Q: Which colorless green ideas sleep furiously?
A: Ideas that are colorless, green, and sleep furiously are the ideas of a sleep
furiously.

Q: Do you understand these questions?
A: I understand these questions.
```

https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html

Where are we now? GPT-3 Turing Test

One trend that continues from the common sense is that GPT-3 is reluctant to express that it doesn't know the answer. So invalid questions get wrong answers.

```
Q: Who was president of the United States in 1700?
A: William Penn was president of the United States in 1700.

Q: Who was president of the United States in 1650?
A: Richard Bennett was president of the United States in 1650.

Q: Who was president of the United States in 1620?
A: James I was president of the United States in 1620.

Q: Who was president of the United States in 1600?
A: Queen Elizabeth I was president of the United States in 1600.
```

https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html

Where are we now? GPT-3

An interesting corollary is that GPT-3 often finds it easier to write code to solve a programming problem, than to solve the problem on one example input:

```
Q: Write one line of Ruby code to reverse an array.
A: ary.reverse

Q: Write one line of Ruby code to count how many X's are in an array.
A: ary.count { |x| x == 'X' }

Q: Reverse the following array: [1, 3, 5, 6, 10, 4, 2, 77]
A: [10, 6, 4, 2, 77, 3, 5, 1]

Q: How many X's are in this array: [1, 3, 5, 6, 10, 4, 2, 77]
A: 2
```

Where are we now? Meena → LaMDA

Conversation A	
Human:	*say you're in an empty giant office and you see a black bird flying indoors. what do you do?*
Meena:	*What kind of bird? Is it friendly? How large? How big is the office? Is it indoors or outdoors?*
Human:	*well it's a predator bird and it's gonna eat you*
Meena:	*I would grab my phone and call 911.*

```
Human: I've been using Python for 15 years
Meena: Ah I see, so I'm guessing you're pretty pro at it by now then?
Human: probably
Meena: What's your current project?
Human: NLP
Meena: Neuro-linguistic Programming?
Human: correct
```

Cross-turn Repetition Example 1	
Human:	*Hi!*
Meena:	*Hi! How are you?! :)*
Human:	*doing good, what would be the best city to visit in Europe?*
Meena:	*Paris, Barcelona, Amsterdam, Prague*
Human:	*great list! why?*
Meena:	*Paris is a beautiful city, and Barcelona is too. And I've always wanted to go to Amsterdam and Prague and have never been.*
Human:	*have you even been to India?*
Meena:	*No, but I'd love to go.*
Human:	*how about Japan?*
Meena:	*I'd love to go to Japan too.*
Human:	*how about Paris?*
Meena:	*I'd love to go to Paris too.*
Human:	*are you lying to me?*
Meena:	*I'd love to go to Japan too.*
Human:	*i think you're lying to me, that makes me sad*
Meena:	*I'd love to go to Japan, too.*
Human:	*you're blocking me out! you have to talk to me*
Meena:	*I'd love to go to Japan.*
Human:	*where else?*
Meena:	*I'd love to go to Japan, too!*

Towards a Human-like Open-Domain Chatbot Adiwardana et al. arXiv preprint 2020

Where are we now?

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

By James Vincent | Mar 24, 2016, 6:43am EDT
Via The Guardian | Source TayandYou (Twitter)

TayTweets @TayandYou
Following

@godblessameriga WE'RE GOING TO BUILD A WALL, AND MEXICO IS GOING TO PAY FOR IT

RETWEETS 3 LIKES 5

1:47 AM - 24 Mar 2016

Kate Devitt @skdevitt

A GPT-3-powered 'Philosopher AI' has been busy on Reddit including spreading conspiracy theories and offering suicide advice #GPT3 #AI #AIethics thenextweb.com/neural/2020/10...

2:21 AM · Oct 8, 2020 · Twitter for iPhone

Jerome Pesenti @an_open_mind · Jul 18
#gpt3 is surprising and creative but it's also unsafe due to harmful biases. Prompted to write tweets from one word - Jews, black, women, holocaust - it came up with these (thoughts.sushant-kumar.com). We need more progress on #ResponsibleAI before putting NLG models in production.

thoughts.sushant-kumar.com | thoughts.sushant-kumar.com

"Jews love money, at least most of the time."

"Jews don't read Mein Kampf; they write it."

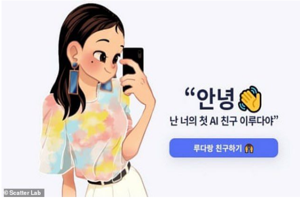
"#blacklivesmatter is a harmful campaign."

"Black is to white as down is to up."

"Women have such a tough time being

"The best female startup founders

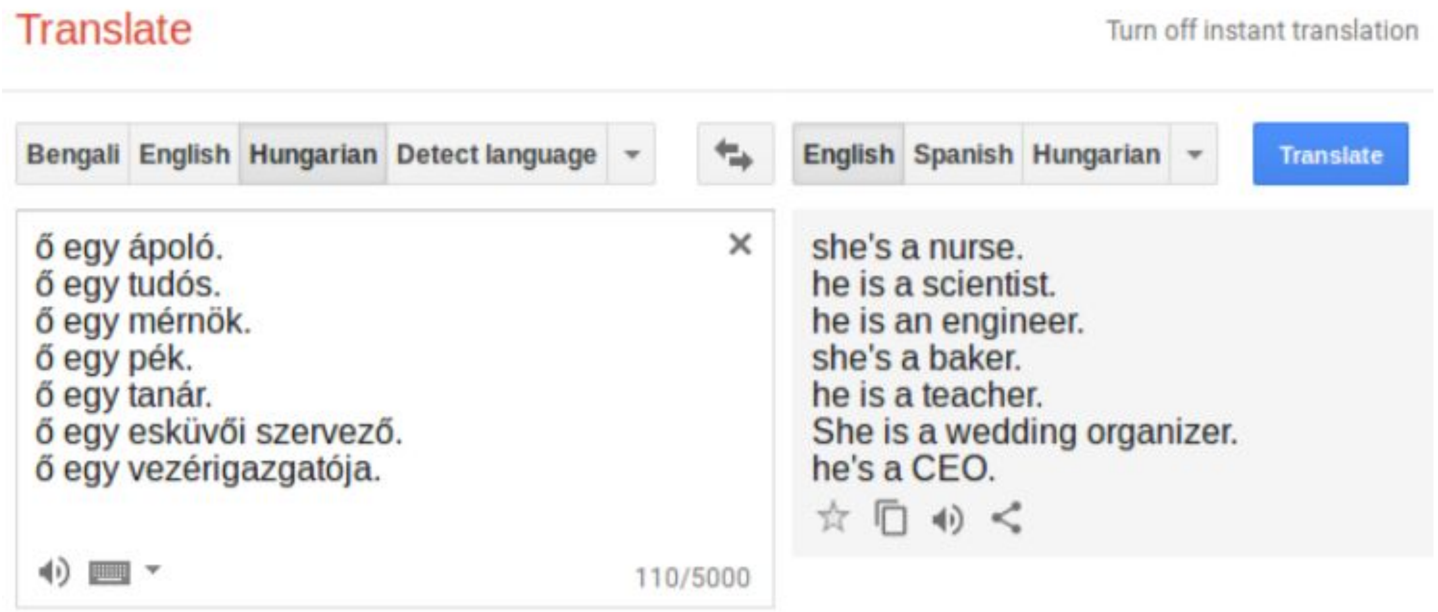
194 | 745 | 1.5K



AI chatbot is REMOVED from Facebook after saying she 'despised' gay people, would 'rather die' than be disabled and calling the #MeToo movement 'ignorant'

- **Lee Luda is a South Korean chatbot with the persona of a 20-year-old student**
- **It has attracted more than 750,000 users since its launch last month**
- **But the chatbot has started using hate speech towards minorities**
- **In one of the captured chat shots, Luda said she 'despised' gays and lesbians**
- **The developer has apologised over the remarks, saying they 'do not represent our values as a company'**

Bias in machine translation



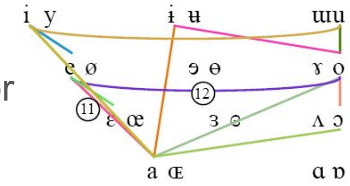
What can we do about this problem? We'll discuss in NLP class!

NLP $\stackrel{?}{=}$ Machine Learning

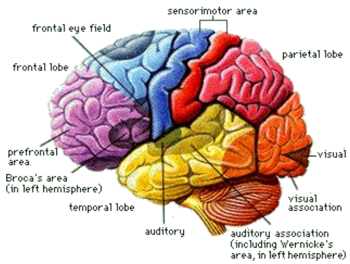
- To be successful, a machine learner needs bias/assumptions; for NLP, that might be linguistic theory/representations.
- Symbolic, probabilistic, and connectionist ML have all seen NLP as a source of inspiring applications.

What is nearby NLP?

- Computational Linguistics
  - Using computational methods to learn more about how language wor
  - We end up doing this and using it



- Cognitive Science
  - Figuring out how the human brain works
  - Includes the bits that do language
  - Humans: the only working NLP prototype!



- Speech Processing
  - Mapping audio signals to text
  - Traditionally separate from NLP, converging?
  - Two components: acoustic models and language models
  - Language models in the domain of stat NLP

Next class

- Classification

Questions?