

Natural Language Processing

Introduction, Text classification

Yulia Tsvetkov

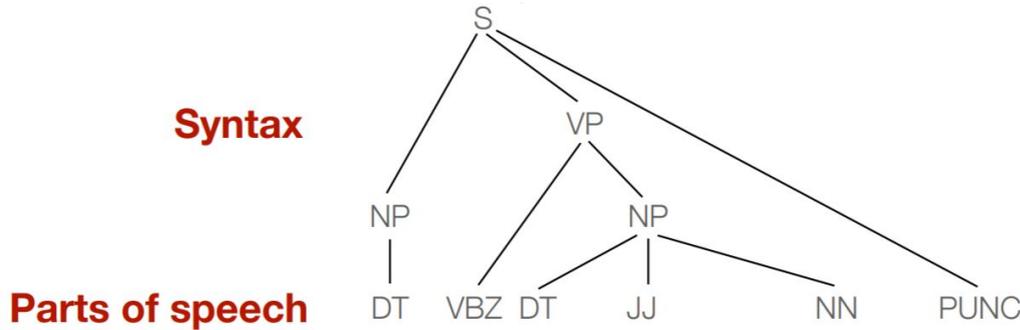
yuliats@cs.washington.edu

Announcements

- Class recordings
- Homework project examples <http://demo.clab.cs.cmu.edu/11711fa20/>

Discourse

- Reference resolution
- Discourse parsing



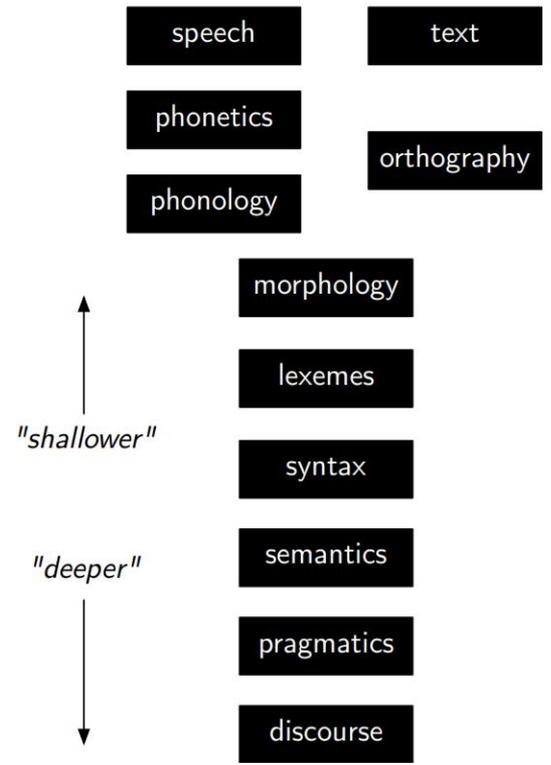
Tokens This is a simple sentence .

Morphology be SIMPLE1: having few parts SENTENCE1: String of words satisfying the grammatical rules of a language

Semantics 3sg present

Discourse But an instructive one .

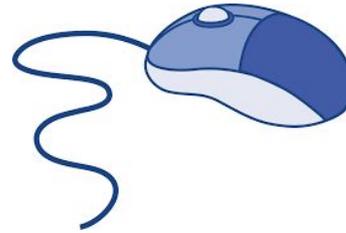
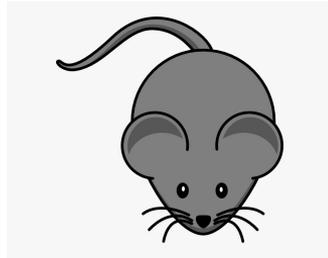
coreferent



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



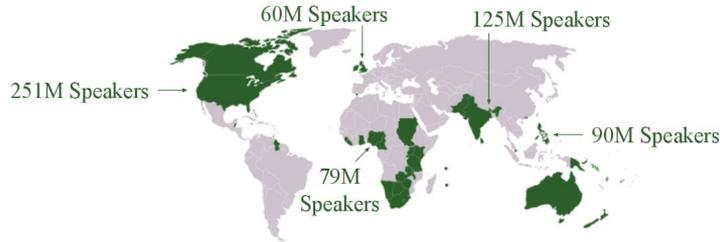
- Russian has very few words for colors, Japanese has hundreds
- Multiword expressions, e.g. [happy as a clam](#), [it's raining cats and dogs](#) or [wake up](#) and metaphors, e.g. [love is a journey](#) are very different across languages

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



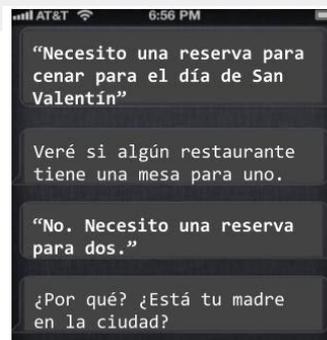
Englishes



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

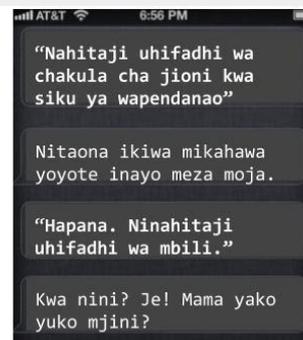
- ~7,000 languages
- thousands of language varieties



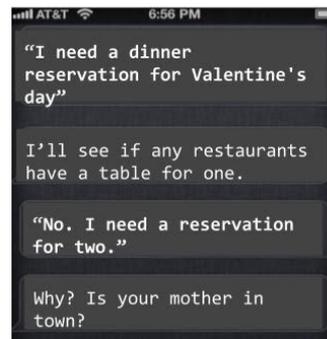
Spanish
534 million speakers



Hindi
615 million speakers



Swahili
100 million speakers



American English



Scottish English

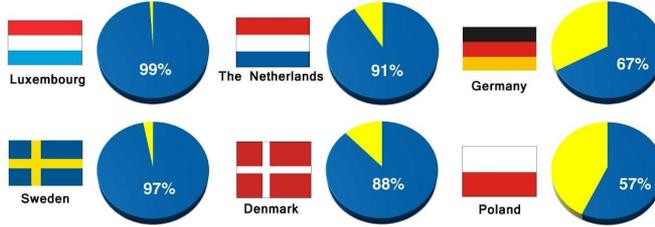


Hinglish

Most of the world today is multilingual

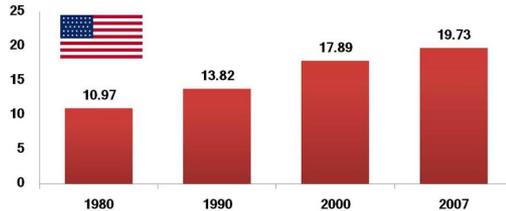
Percentage of Bilingual Speakers in the World

European Union



Source: European Commission, "Europeans and their Languages," 2006

Percentage of US Population who spoke a language other than English at home by year

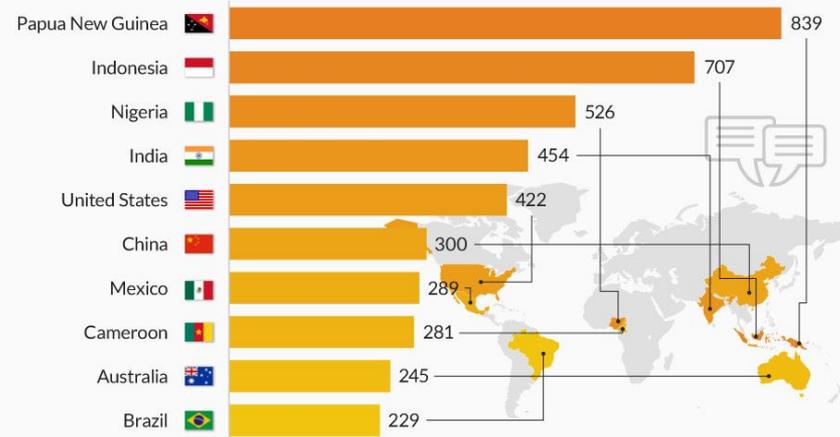


Source: U.S. Census Bureau, 2007 American Community Survey

Source: US Census Bureau

The Countries With The Most Spoken Languages

Number of living languages spoken per country in 2015



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
her daughter

בתה

- 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'ananyakapushasqakupuniñataqsunamá

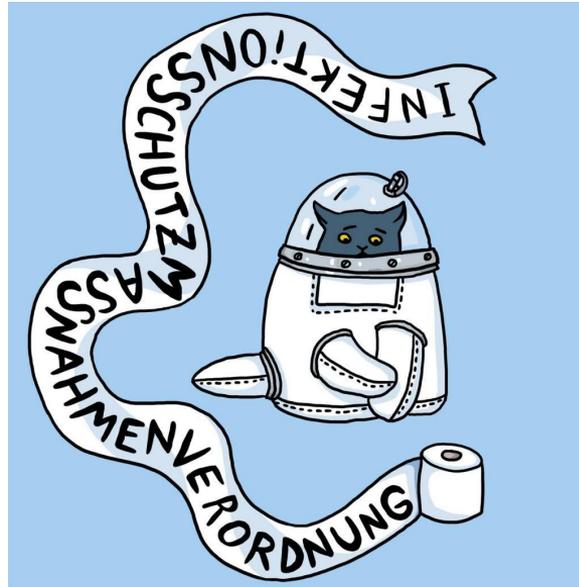
Much'a -na -naya -ka -pu -sha -sqa -ku -puni -ña -taq -suna -má

"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

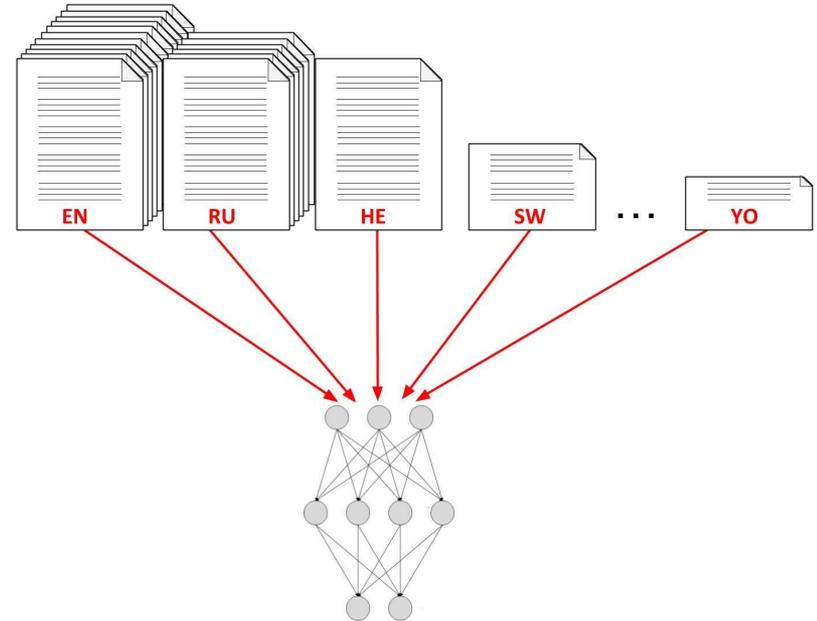
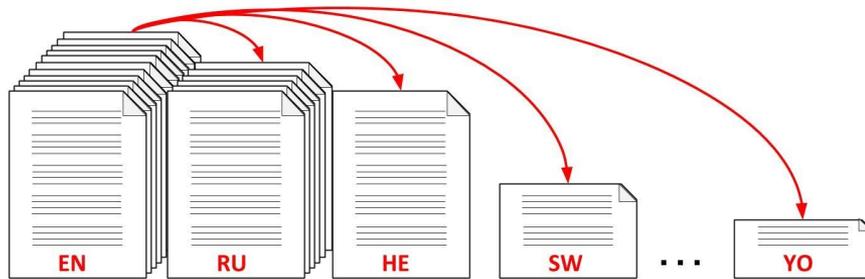
- German



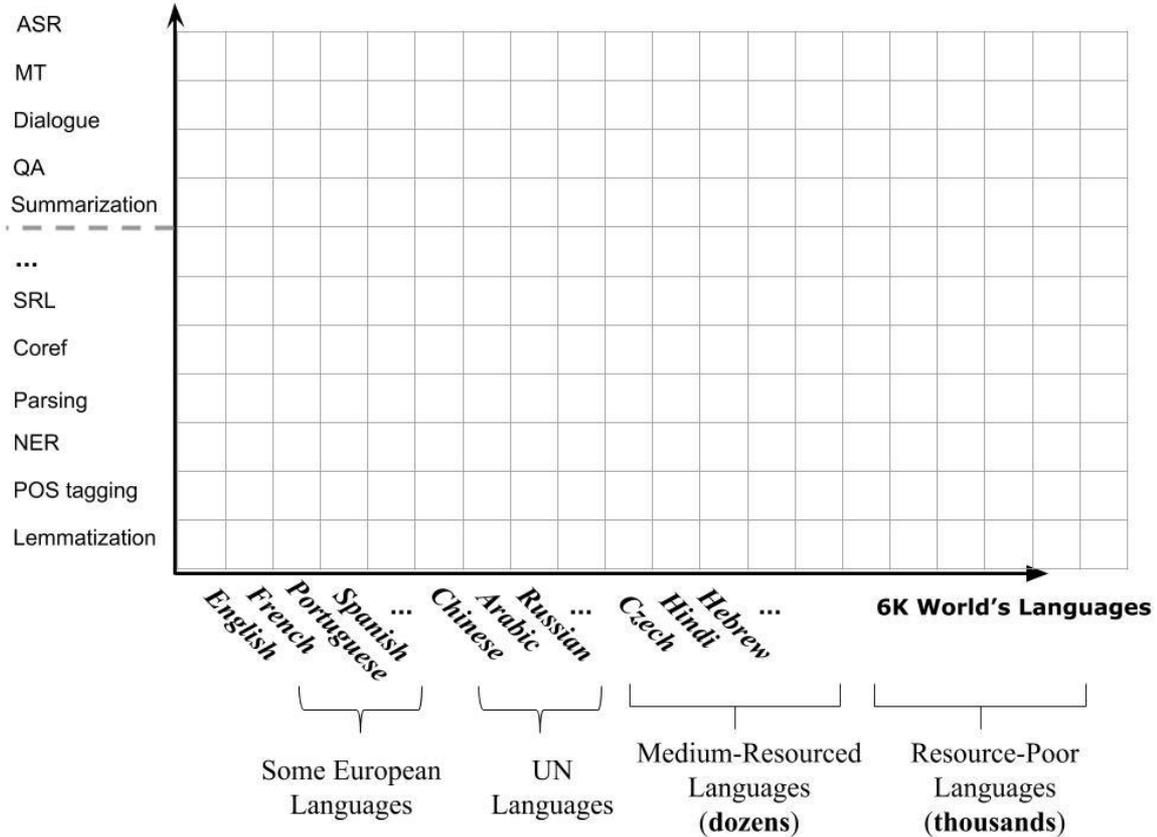
Infektionsschutzmaßnahmenverordnung

Multilingual NLP

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



NLP Technologies/Applications



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

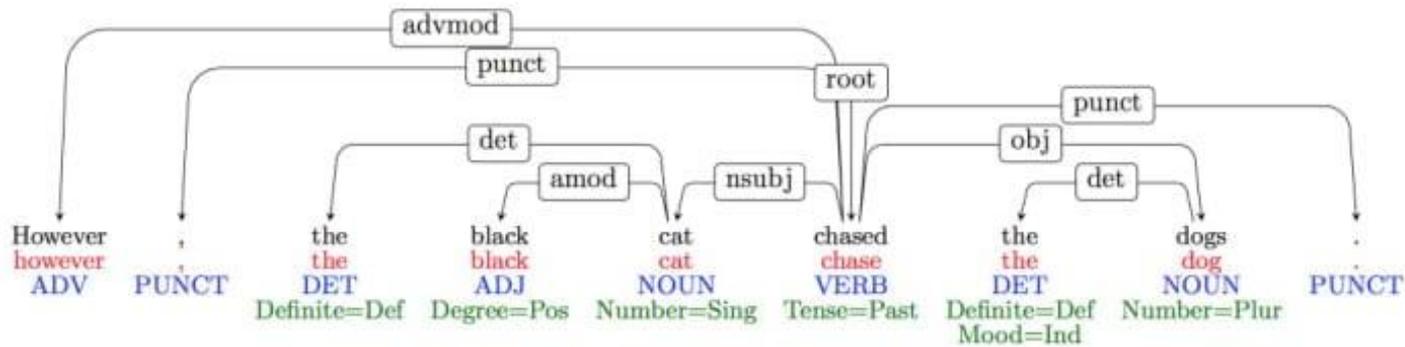
- Non-standard language, emojis, hashtags, names



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

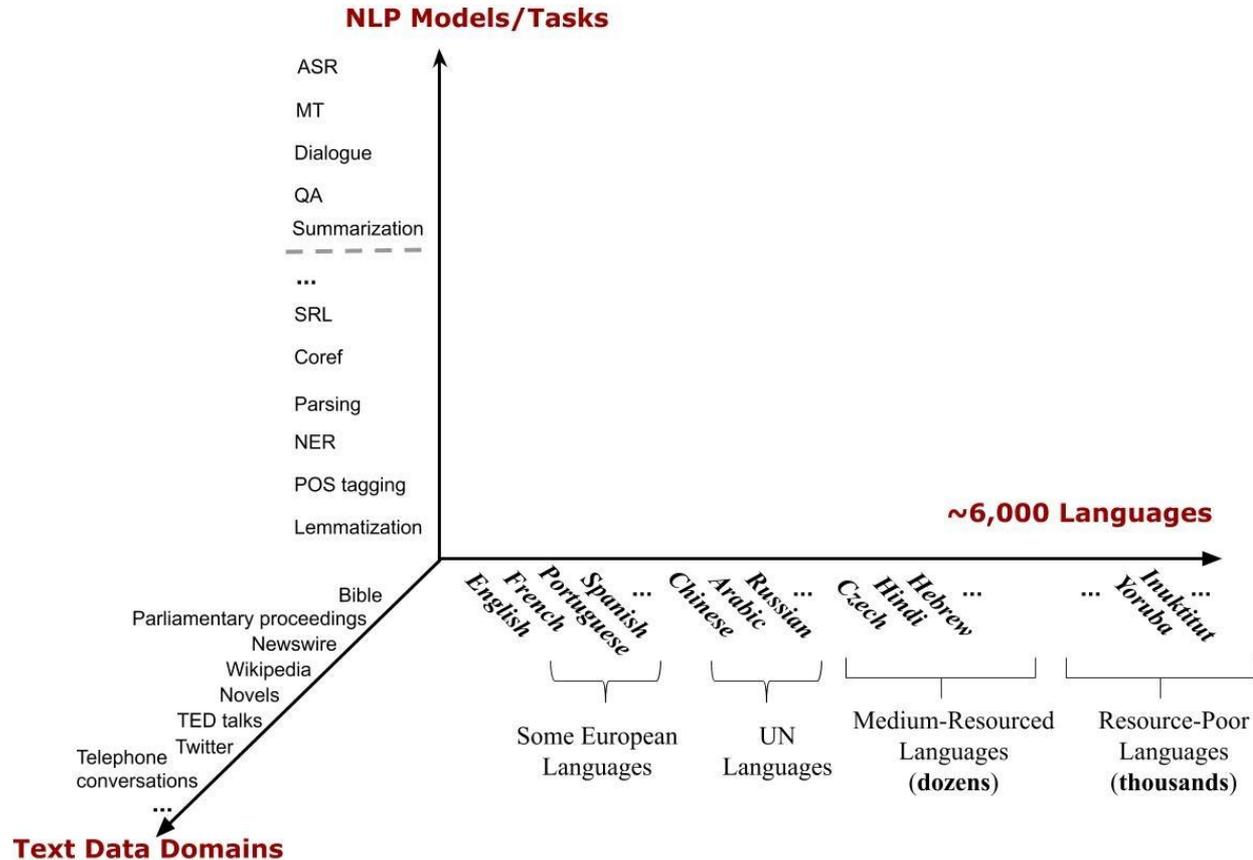
Variation

- 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??

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



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

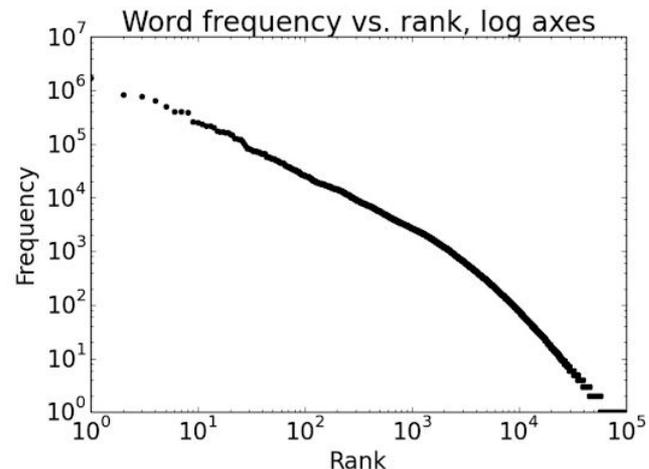
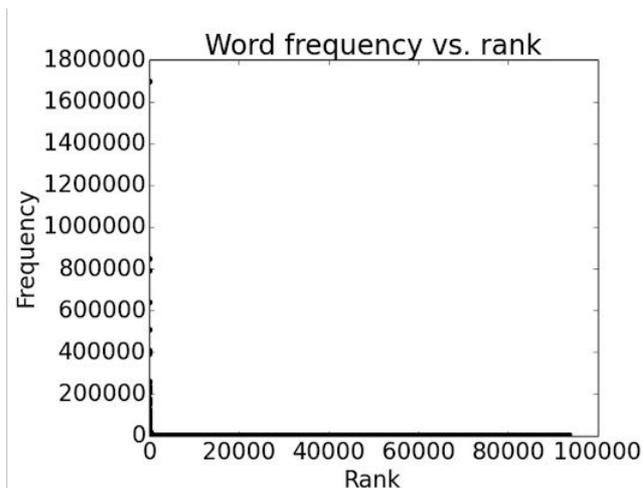
But also, out of 93,638 distinct words (word types), 36,231 occur only once.

Examples:

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

Plotting word frequencies

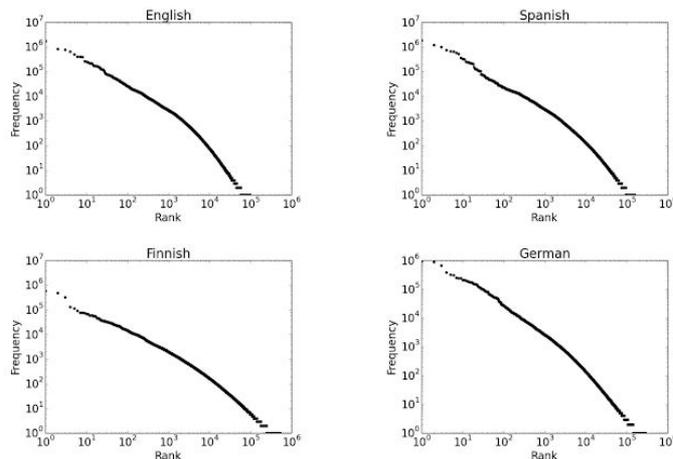
Order words by frequency. What is the frequency of n th ranked word?



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

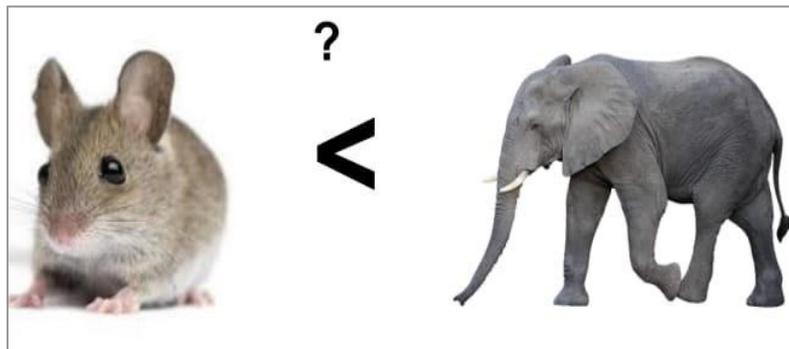
Why is language interpretation hard?

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?

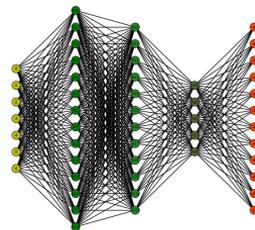
1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. Unknown representation \mathcal{R}

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

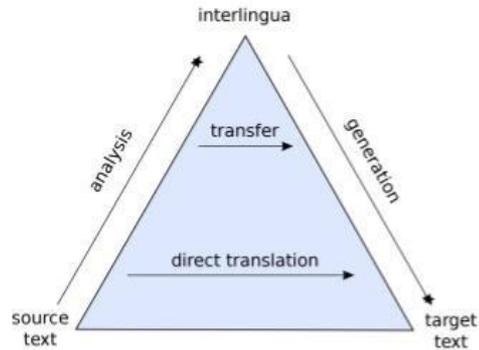


Desiderata for NLP models

- 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

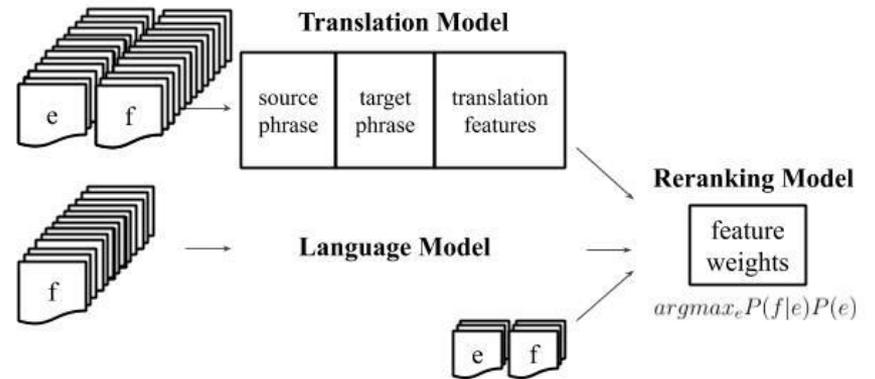
Logic-based/Rule-based NLP



~ 90s

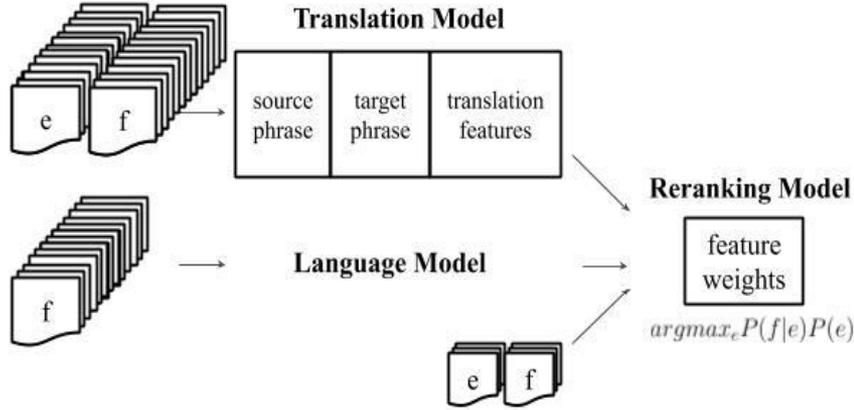


Statistical NLP



Probabilistic and Connectionist NLP

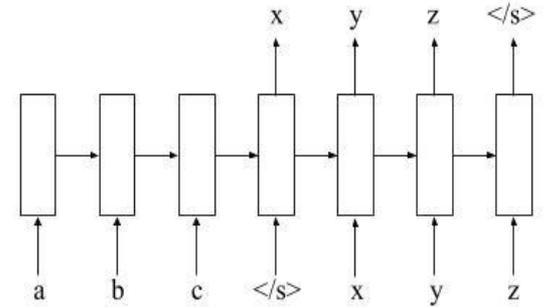
Engineered Features/Representations



~mid 2010s



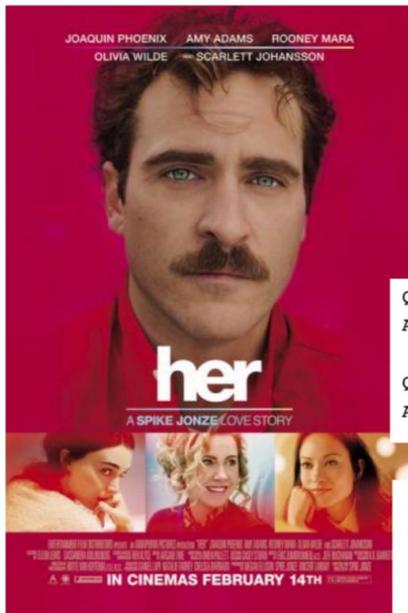
Learned Features/Representations



GPT-3



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.*

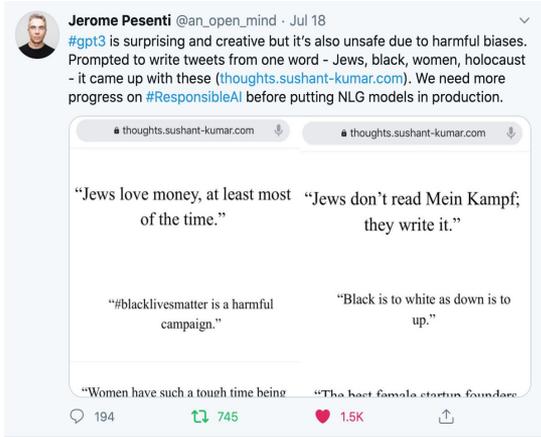
Human: *where else?*

Meena: *I'd love to go to Japan, too!*

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)



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

Translate

Turn off instant translation

The screenshot shows the Google Translate interface. The source language is set to Hungarian and the target language is English. The input text is a list of Hungarian sentences, each starting with "Ő egy" (He/She is a). The output text shows the machine's translations, which are biased towards male and professional roles.

Original Hungarian	Machine Translated English
Ő egy ápoló.	she's a nurse.
Ő egy tudós.	he is a scientist.
Ő egy mérnök.	he is an engineer.
Ő egy pék.	she's a baker.
Ő egy tanár.	he is a teacher.
Ő egy esküvői szervező.	She is a wedding organizer.
Ő egy vezérigazgatója.	he's a CEO.

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

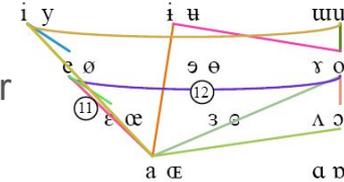
NLP $\stackrel{?}{=}$ Machine Learning

- 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?

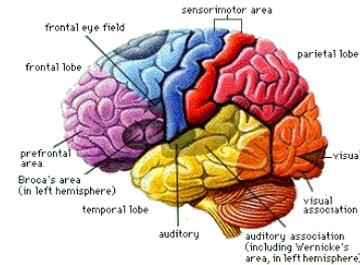
- Computational Linguistics

- Using computational methods to learn more about how language works
- 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 topic

- Classification

Questions?

Is this spam?

from: **ECRES 2022 <2022@ecres.net>** [via](#) amazonses.com
reply-to: 2022@ecres.net
to: yuliats@cs.washington.edu
date: Feb 22, 2022, 7:21 AM
subject: The Best Renewable Energy Conference (Last chance !)
signed-by: amazonses.com
security: Standard encryption (TLS) [Learn more](#)

Dear Colleague,

Account: yuliats@cs.washington.edu

Good news: Due to many requests, the submission deadline has been extended to **10 March 2022** (It is firm date).

We would like to invite you to submit a paper to 10. European Conference on Renewable Energy Systems (ECRES). **ECRES 2022 will be held hybrid mode, the participants can present their papers physically or online.** The event is going to be organized in Istanbul/Turkey under the technical sponsorship of Istanbul Medeniyet University and many international institutions. The conference is highly international with the participants from all continents and more than 40 countries.

The submission deadline and special and regular issue journals can be seen in [ecres.net](#)

There will be keynote speakers who will address specific topics of energy as you would see at [ecres.net/keynotes.html](#)

[CLICK FOR PAPER SUBMISSION](#)

All accepted papers will be published in a special Conference Proceedings under a specific ISBN. Besides, the extended versions will be delivered to reputable journals **indexed in SCI, E-SCI, SCOPUS, and EBSCO**. You can check our previous journal publications from [ecres.net](#) . **Please note that the official journal of the event, Journal of Energy Systems ([dergipark.org.tr/jes](#)) is also indexed in SCOPUS.**

Spam classification

Dear Colleague,

Account: yuliats@cs.washington.edu

Good news: Due to many requests, the submission deadline has been extended to 10 March 2022 (It is firm date).

We would like to invite you to submit a paper to the conference on Renewable Energy Systems (ECRES). **ECRES 2022** will be held in Istanbul/Turkey under the technical sponsorship of Medeniyet University and many international institutions. The conference is international with the participants from all continents and more than 40 countries.



The submission deadline and special and regular issue journals can be seen in ecres.net

There will be keynote speakers who will address specific topics of energy as you would see at ecres.net/keynotes.html

[CLICK FOR PAPER SUBMISSION](#)

All accepted papers will be published in a special Conference Proceedings under a specific ISBN. Besides, the extended versions will be delivered to reputable journals indexed in **SCI, E-SCI, SCOPUS, and EBSCO**. You can check our previous journal publications from ecres.net. **Please note that the official journal of the event, Journal of Energy Systems (dergipark.org.tr/jes) is also indexed in SCOPUS.**

Invitation to present at the February 2022 Wikimedia Research Showcase



Emily Lescak <elescak@wikimedia.org>
to yuliats@cs.washington.edu

Hi Yulia,

My name is Emily Lescak and I am a member of the [Research Team](#) at the Wikimedia Foundation. On behalf of the Research Team, I would like to invite you to present your research on social biases on Wikipedia at our [Research Showcase](#) in February 2022. This topic fits into our theme for this showcase, which is gaps and biases on Wikipedia.

The Wikimedia Research Showcase is a monthly, public lecture series where Foundation, academic, and community members present their work related to Wikipedia, Wikimedia, peer production, and open-source software. We focus on topics and projects that we think our audience—a global community of academic researchers, Wikipedia editors, and Wikimedia staff—would find interesting and relevant to their work.

Research Showcase presentations are generally 20 minutes long, with an additional 10 minutes for questions. We invite two presenters to every showcase. Most presenters choose to use slides to present their work.

The February showcase takes place on the 16th at 9:15AM Pacific / 17:15 UTC. You can watch past showcases on our [YouTube](#) and also archived for later viewing on the [Wikimedia Foundation's YouTube channel](#)

If this date does not work for you, but you are still interested in giving a showcase, please let us know so we can discuss other options.

I hope to get a chance to see your work presented at the Research Showcase!

Sincerely,

Emily

--



Language ID

Аяны замд түр зогсон тэнгэрийн байдлыг ажиглаад хөдлөх зуур гутал дор шинэхэн орсон цас шаржигнан дуугарч байв. Цасны тухай бодол сонин юм. Хот хүрээ тийш цас орвол орно л биз гэсэн хэнэггүй бодол маань хөдөө талд, говийн ээрэм хөндийд, малын бэлчээрт, малчдын хотонд болохоор солигдож эргэцүүлэн бодох нь хачин. Цас хэр орсон бол?

Београд, 16. јун 2013. године – Председник Владе Републике Србије Ивица Дачић честитао је кајакашици златне медаље у олимпијској дисциплини К-1, 500 метара, као и у двоструко дужој стази освојене на првенству Европе у Португалији.

Beograd, 16. јun 2013. године – Председник Владе Републике Србије Ivica Dačić čestitao је кајакашиси златне medalје u оlimpiјској disciplini K-1, 500 metara, као i u dvostruko dužој stazi освојене на првенstvu Evроpe u Portugaliji.

Nestranski Urad за vladno odgovornost ZDA је objavil eksplozivno mnenje, da је vlada predsednika Donalda Trampa kršila zvezno zakonodajo, ko је zadrževala izplačilo kongresno potrjene vojaške pomoči Ukrajini zaradi političnih razlogov. Predstavniški dom kongresa је prav zaradi tega sprožil ustavno obtožbo proti Trumpu.

Language ID

Аяны замд түр зогсон тэнгэрийн байдлыг ажиглаад хөдлөх зуур гутал дор шинэхэн орсон цас шаржигнан дуугарч байв. Цасны тухай бодол сонин юм. Хот **mongolian** рвол орно л биз гэсэн хэнэггүй бодол маань хөдөө тал **mongolian** өндийд, малын бэлчээрт, малчдын хотонд болохоор солигдож эргэцүүлэн бодох нь хачин. Цас хэр орсон бол?

Београд, 16. јун 2013. године – Председник Владе Републике Срб **serbian** неститао је кајакашици златне медаље у оли **serbian** ини К-1, 500 метара, као и у двоструко дужој стази освојене на првенству Европе у Португалији.

Beograd, 16. јun 2013. године – Председник Владе Републике Ср **bosnian** титао је кајакашици златне медаље у о **bosnian** К-1, 500 метара, као и у двоstrуко дуђој стази освојене на првенству Европе у Португалији.

Nestranski Urad za vladno odgovornost ZDA je objavil eksplozivno mnenje, da je vlada predsednika Donalda Trumpa kršila zvezno zakonodajo, ko je zadrževala izplačilo k **slovenian** vojaške pomoči Ukrajini zaradi političnih razlogov. Predstavniški d **slovenian** v zaradi tega sprožil ustavno obtožbo proti Trumpu.

Sentiment analysis



By [John Neal](#)

This review is from: [Accoutrements Horse Head Mask \(Toy\)](#)

When I turned State's Witness, they didn't have enough money to put me in the Witness Protection Program, so they bought me this mask and gave me a list of suggested places to move. Since then I've lived my life in peace and safety knowing that my old identity is forever obscured by this life-saving item.



By [Christine E. Torok](#)

Verified Purchase ([What's this?](#))

First of all, for taste I would rate these a 5. So good. Soft, true-to-taste fruit flavors like the sugar variety...I was a happy camper.

BUT (or should I say BUTT), not long after eating about 20 of these all hell broke loose. I had a gastrointestinal experience like nothing I've ever imagined. Cramps, sweating, bloating beyond my worst nightmare. I've had food poisoning from some bad shellfish and that was almost like a skip in the park compared to what was going on inside

Sentiment analysis



By [John Neal](#)

This review is from: [Accoutrements Horse Head Mask \(Toy\)](#)

When I turned State's Witness, they didn't have enough money to put me in the Witness Protection Program, so they bought me this mask and I came all over sugar and places to move. Since then, I've lived my life in peace and safety knowing that my old identity is forever obscured by this life-saving item.



By [Christine E. Torok](#)

Verified Purchase ([What's this?](#))

First of all, for taste I would rate these a 5. So good. Soft, true-to-taste fruit flavors like the sugar variety...I was a happy camper.

BUT (or should I say BUTT), not long after eating about 20 of these all hell broke loose. I had a gastrointestinal experience like nothing I've ever imagined. Cramps, sweating, floating beyond my worst nightmare. I've had food poisoning from some bad shellfish and that was almost like a skip in the park compared to what was going on inside



Topic classification

MEDLINE Article

The image shows a snippet of a MEDLINE article. At the top, it features the Elsevier logo and the journal title 'Brain Cognition'. The title of the article is 'Syntactic frame and verb bias in aphasia: Plausibility judgments of underdog-subject sentences'. Below the title, the authors are listed: Susanna Gulló, Lisa Mann, Gail Ramscar, David R. Jansky, Elizabeth Riley, Moly Ravegg, and L. Holland Aulrey. The article is published in 'Brain Cognition', volume 152, pages 1-12, in 2020. The abstract begins with 'The study investigates how native fluents have acquired to derive "underdog" subject sentences...'. The main text starts with '1. Introduction' and discusses the complexity of "underdog" subject sentences and how they are processed by the brain.



MeSH Subject Category Hierarchy

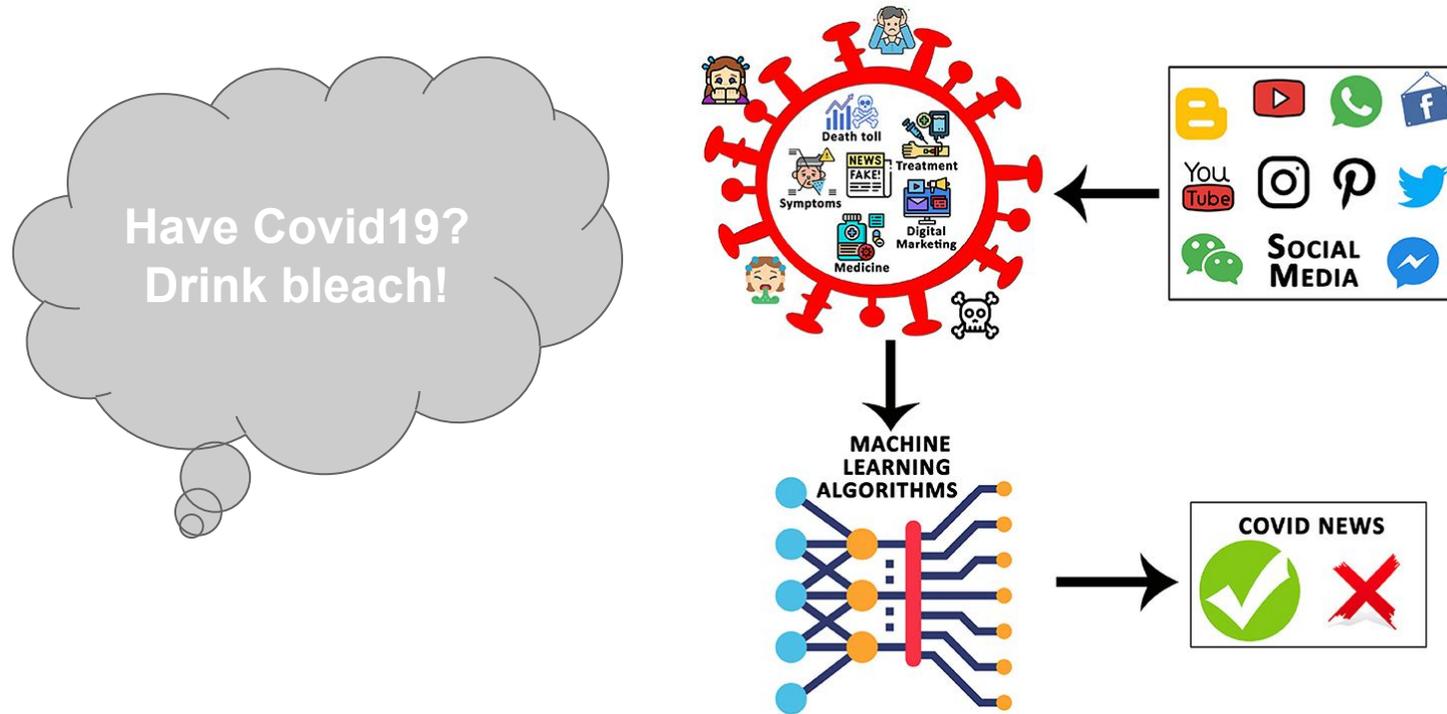
- Antagonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- ...

Authorship attribution: is the author male or female?

By 1925 Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony Cochinchina; the central area with its imperial capital at Hue was the protectorate of Annam.

Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of the greatest assets...

Fact verification: trustworthy or fake?



Detecting COVID-19-Related Fake News Using Feature Extraction

Suleman Khan, Saqib Hakak, N. Deepa, B. Prabadevi, Kapal Dev and Silvia Trelova

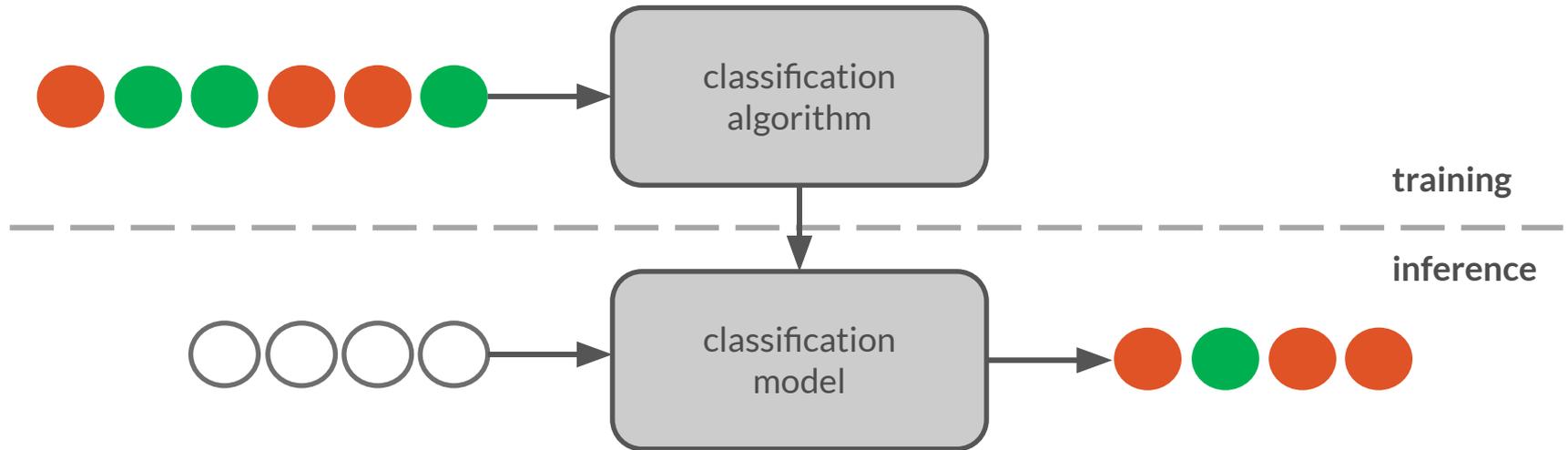
Text classification

- We might want to categorize the content of the text:
 - Spam detection (binary classification: spam/not spam)
 - Sentiment analysis (binary or multiway)
 - movie, restaurant, product reviews (pos/neg, or 1-5 stars)
 - political argument (pro/con, or pro/con/neutral)
 - Topic classification (multiway: sport/finance/travel/etc)
 - Language Identification (multiway: languages, language families)
 - ...
- Or we might want to categorize the author of the text (authorship attribution)
 - Human- or machine generated?
 - Native language identification (e.g., to tailor language tutoring)
 - Diagnosis of disease (psychiatric or cognitive impairments)
 - Identification of gender, dialect, educational background, political orientation (e.g., in forensics [legal matters], advertising/marketing, campaigning, disinformation)
 - ...

Classification: learning from data

- Supervised
 - labeled examples
 - Binary (true, false)
 - Multi-class classification (politics, sports, gossip)
 - Multi-label classification (#party #FRIDAY #fail)
- Unsupervised
 - no labeled examples
- Semi-supervised
 - labeled examples + non-labeled examples
- Weakly supervised
 - heuristically-labeled examples

Supervised classification



Classification: features (measurements)

- Perform measurements and obtain features



4.2, 212, 3.4, 1332
↓ ↓ ↓ ↓
diameter, weight, softness, color



5.2, 315, 5.7, 4567
↓ ↓ ↓ ↓
diameter, weight, softness, color

Supervised classification: formal setting

- Learn a **classification model** from labeled data on
 - properties (“features”) and their importance (“weights”)
- **X**: set of attributes or features $\{x_1, x_2, \dots, x_n\}$
 - e.g. fruit measurements, or word counts extracted from an input documents
- **y**: a “class” label from the label set $Y = \{y_1, y_2, \dots, y_k\}$
 - e.g., fruit type, or spam/not spam, positive/negative/neutral

Supervised classification: formal setting

- Learn a **classification model** from labeled data on
 - properties (“features”) and their importance (“weights”)
- **X**: set of attributes or features $\{x_1, x_2, \dots, x_n\}$
 - e.g. fruit measurements, or word counts extracted from an input documents
- **y**: a “class” label from the label set $Y = \{y_1, y_2, \dots, y_k\}$
 - e.g., fruit type, or spam/not spam, positive/negative/neutral

- Given data samples $\{x_1, x_2, \dots, x_n\}$ and corresponding labels $Y = \{y_1, y_2, \dots, y_k\}$
- We **train** a function $f: x \in X \rightarrow y \in Y$ (the model)

Supervised classification: formal setting

- Learn a **classification model** from labeled data on
 - properties (“features”) and their importance (“weights”)
- **X**: set of attributes or features $\{x_1, x_2, \dots, x_n\}$
 - e.g. fruit measurements, or word counts extracted from an input documents
- **y**: a “class” label from the label set $Y = \{y_1, y_2, \dots, y_k\}$
 - e.g., fruit type, or spam/not spam, positive/negative/neutral

- At **inference** time, apply the model on new instances to **predict the label**

Text classification – feature extraction

What can we measure over text? Consider this movie review:

I love this movie! It's sweet, but with satirical humor. The dialogue is great, and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it just to about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it before.

Text classification – feature extraction

What can we measure over text? Consider this movie review:

I **love** this movie! It's **sweet**, but with **satirical humor**. The dialogue is **great**, and the adventure scenes are **fun**... It manages to be **whimsical** and **romantic** while **laughing** at the conventions of the fairy tale genre. I would **recommend** it just to about anyone. I've seen it **several** times, and I'm always happy to see it **again** whenever I have a friend who hasn't seen it before.

Text classification – feature extraction

I **love** this movie! It's **sweet**, but with **satirical humor**. The dialogue is **great**, and the adventure scenes are **fun**... It manages to be **whimsical** and **romantic** while **laughing** at the conventions of the fairy tale genre. I would **recommend** it just to about anyone. I've seen it **several** times, and I'm always happy to see it **again** whenever I have a friend who hasn't seen it before.

(almost) the entire lexicon

word	count	relative frequency
love	10	0.0007
great	...	
recommend		
laugh		
happy		
...		
several		
boring		
...		

Types of textual features

- Words
 - content words, stop-words
 - punctuation? tokenization? lemmatization? lowercase?
- Word sequences
 - bigrams, trigrams, n-grams
- Grammatical structure, sentence parse tree
- Words' part-of-speech
- Word vectors
- ...

We'll consider alternative models for classification

