



Language
Technologies
Institute

Carnegie
Mellon
University

Algorithms for NLP

CS 11711, Fall 2019

Lecture 1: Introduction

Yulia Tsvetkov

Welcome!



Yulia



Bob



Sachin



Anjalie



Chan

Course Website

<http://demo.clab.cs.cmu.edu/11711fa19/>

Algorithms for NLP

CMU CS 11711, Fall 2019

T/Th 1:30-2:50pm, Doherty Hall 1212

[Yulia Tsvetkov](#) (office hours: TBD, GHC 6405), ytsvetko@cs.cmu.edu

[Robert Frederking](#) (office hours: TBD), ref@cs.cmu.edu

Teaching Assistants:

[Anjalie Field](#) (office hours: TBD), anjalief@cs.cmu.edu

[Chan Young Park](#) (office hours: TBD), chanyoun@cs.cmu.edu

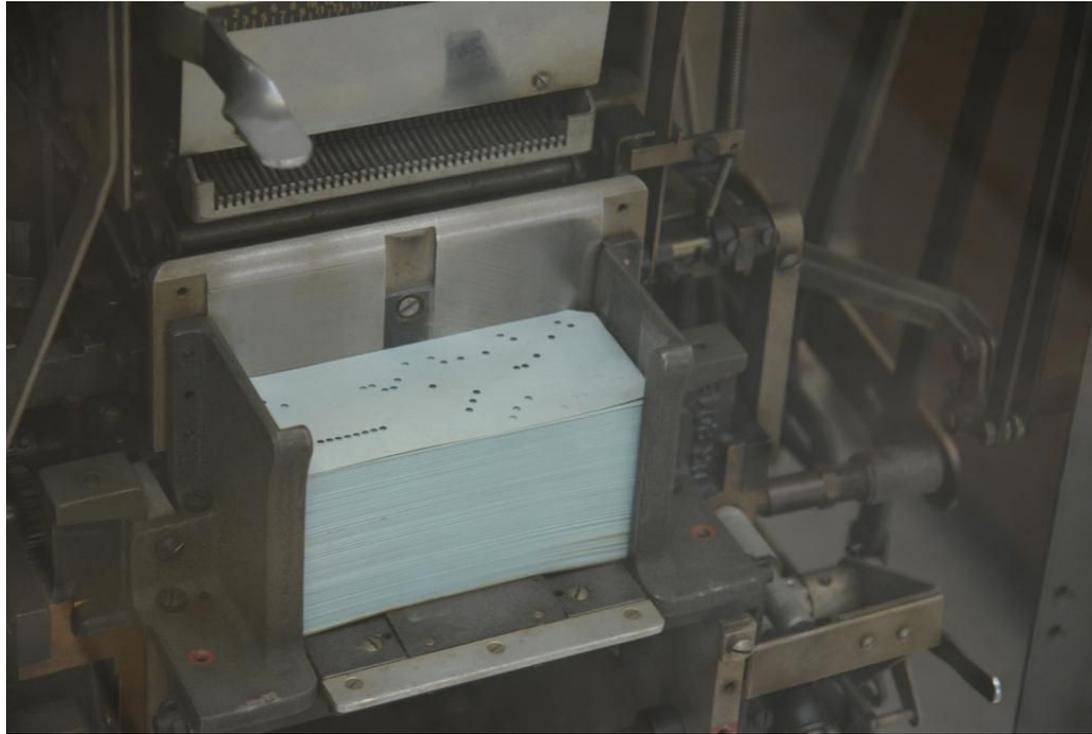
[Sachin Kumar](#) (office hours: TBD), sachink@cs.cmu.edu

Forum: [Piazza](#)



Communication with Machines

- ~50s-70s

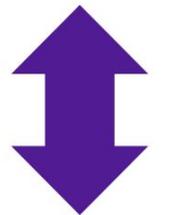


Communication with Machines

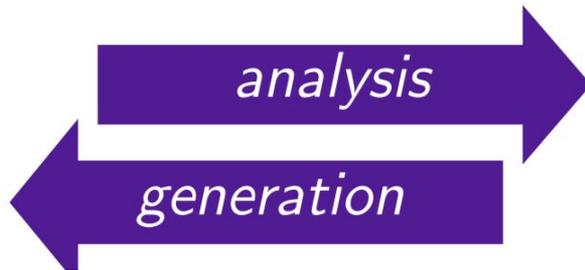
- Today



Slide by Nc



NL



\mathcal{R}

Slide by Noah Smith



What is NLP?

- $NL \in \{\text{Mandarin, Hindi, Spanish, Arabic, English, ... Inuktitut}\}$
- Automation of NLPs:
 - analysis ($NL \rightarrow \mathcal{R}$)
 - generation ($\mathcal{R} \rightarrow NL$)
 - acquisition of \mathcal{R} from knowledge and data



What language technologies are required to write such a program?



Language Technologies



A conversational agent contains

- Speech recognition
- Language analysis
- Dialog processing
- Information retrieval
- Text to speech



Language Technologies



Text and Web

Translated Search

Dictionary

Tools

Translate Text

Original text:

Istotą instytucji wyłączenia organu podatkowego od załatwienia sprawy dotyczącej zobowiązania podatkowego lub innej sprawy normowanej przepisami prawa podatkowego jest utrata właściwości danego organu do załatwienia danej sprawy.

Translation: Polish (automatically detected) »
Finnish

Pelkät vapautusta veron käsittelylle viranomaiselle tapauksissa, joissa verovelan tai muita aineita, normowanej vero-oikeuden menetys kiinteistöä kyseisen viranomaisen ratkaista asian erityinen veronmaksajille.

Detect language »

Finnish »

Translate

[Suggest a better translation](#)

Language

English

Telugu

Swahili

Translate

Detect language	Corsican	Gujarati	Kazakh	Marathi	Shona	Urdu
Afrikaans	Croatian	Haitian Creole	Khmer	Mongolian	Sindhi	Uzbek
Albanian	Czech	Hausa	Korean	Myanmar (Burmese)	Sinhala	Vietnamese
Amharic	Danish	Hawaiian	Kurdish (Kurmanji)	Nepali	Slovak	Welsh
Arabic	Dutch	Hebrew	Kyrgyz	Norwegian	Slovenian	Xhosa
Armenian	English	Hindi	Lao	Pashto	Somali	Yiddish
Azerbaijani	Esperanto	Hmong	Latin	Persian	Spanish	Yoruba
Basque	Estonian	Hungarian	Latvian	Polish	Sundanese	Zulu
Belarusian	Filipino	Icelandic	Lithuanian	Portuguese	Swahili	
Bengali	Finnish	Igbo	Luxembourgish	Punjabi	Swedish	
Bosnian	French	Indonesian	Macedonian	Romanian	Tajik	
Bulgarian	Frisian	Irish	Malagasy	Russian	Tamil	
Catalan	Galician	Italian	Malay	Samoan	Telugu	
Cebuano	Georgian	Japanese	Malayalam	Scots Gaelic	Thai	
Chichewa	German	Javanese	Maltese	Serbian	Turkish	
Chinese	Greek	Kannada	Maori	Sesotho	Ukrainian	



Language Technologies



- What does “divergent” mean?
- What year was Abraham Lincoln born?
- How many states were in the United States that year?
- How much Chinese silk was exported to England in the end of the 18th century?
- What do scientists think about the ethics of human cloning?



NLP

- Applications

- Machine Translation
- Information Retrieval
- Question Answering
- Dialogue Systems
- Information Extraction
- Summarization
- Sentiment Analysis
- ...

- Core technologies

- Language modelling
- Part-of-speech tagging
- Syntactic parsing
- Named-entity recognition
- Coreference resolution
- Word sense disambiguation
- Semantic Role Labelling
- ...



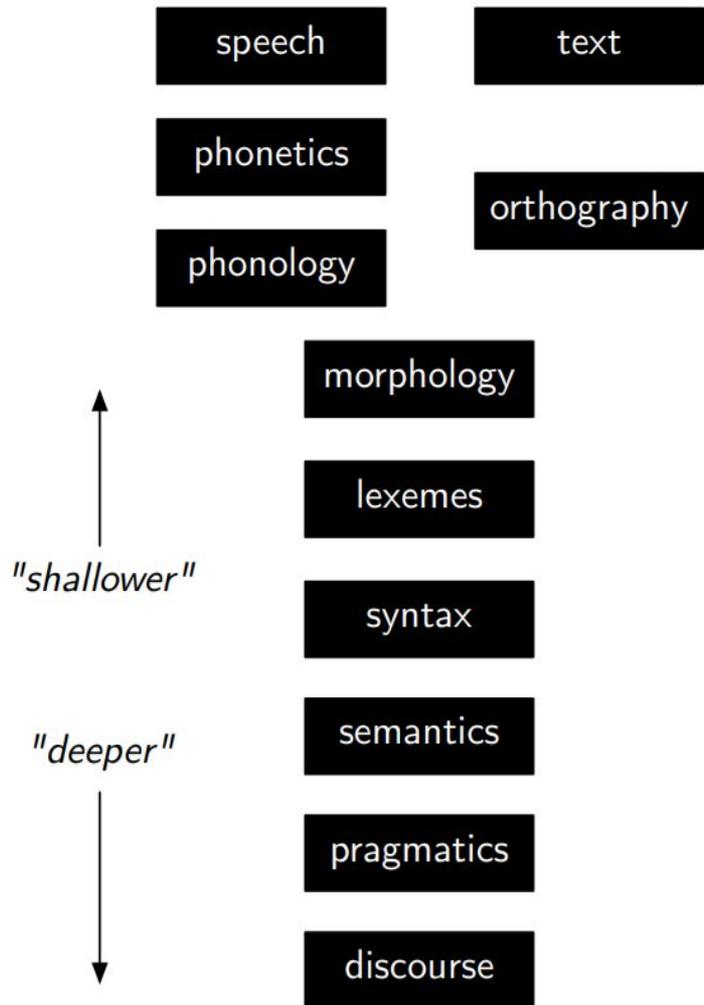
What does an NLP system need to 'know'?

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



What does it mean to “know” a language?

Levels of linguistic knowledge



Phonetics, phonology

- Pronunciation modeling

SOUNDS

Th i a si e n



Words

- Language modeling
- Tokenization
- Spelling correction

WORDS

This is a simple sentence



Morphology

- Morphological analysis
- Tokenization
- Lemmatization

WORDS
MORPHOLOGY

This is a simple sentence

be
3sg
present



Parts of speech

- Part-of-speech tagging

PART OF SPEECH

WORDS

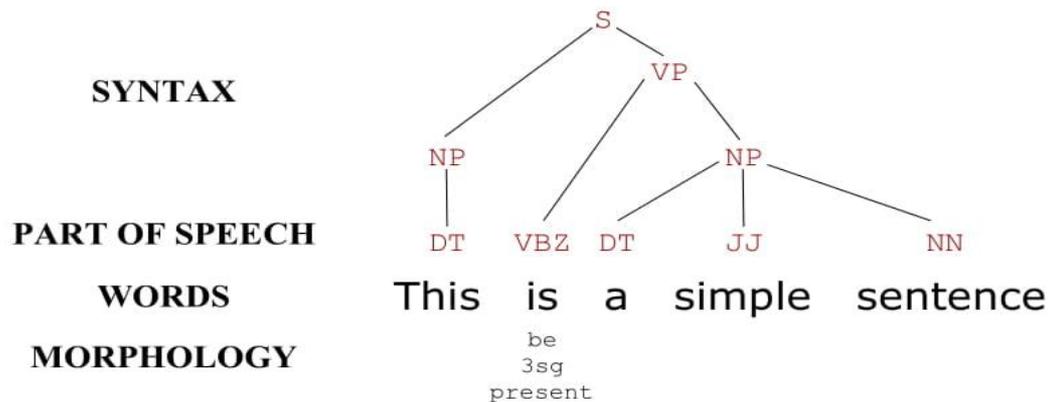
MORPHOLOGY

DT	VBZ	DT	JJ	NN
This	is	a	simple	sentence
	be			
	3sg			
	present			



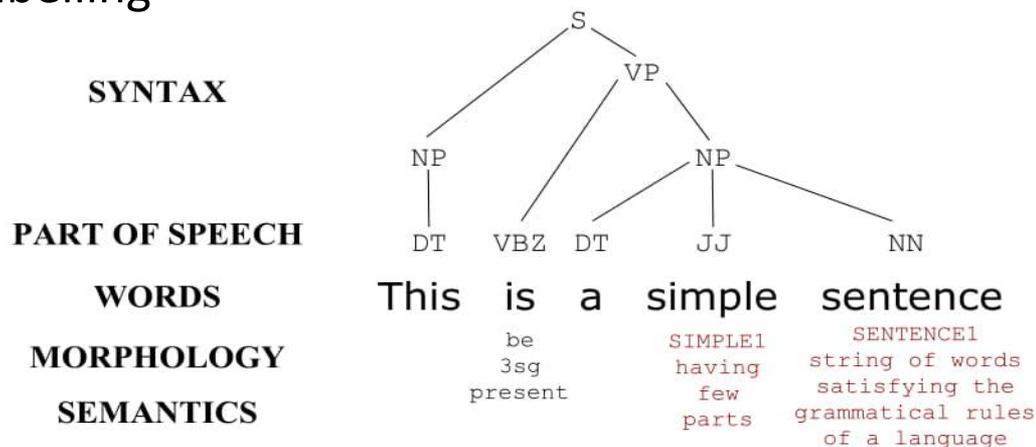
Syntax

- Syntactic parsing



Semantics

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



Discourse

- Reference resolution
- Discourse parsing

SYNTAX

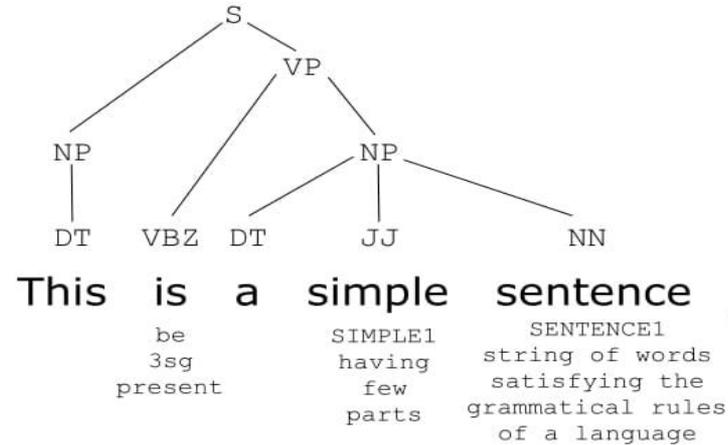
PART OF SPEECH

WORDS

MORPHOLOGY

SEMANTICS

DISCOURSE

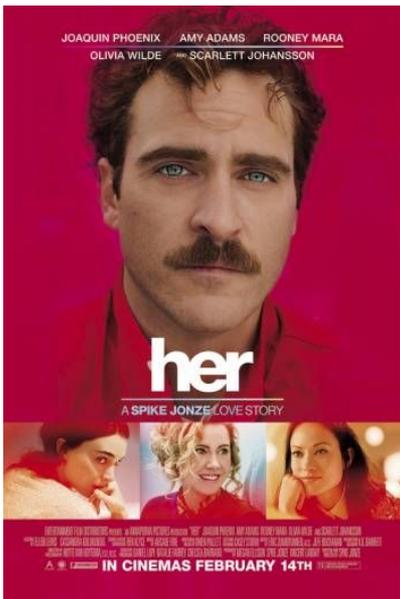


But it is an instructive one.

CONTRAST



Where are we now?



Baseline mutual information model (Li et al. 2015)

A: Where are you going? (1)

B: I'm going to the restroom. (2)

A: See you later. (3)

B: See you later. (4)

A: See you later. (5)

B: See you later. (6)

...

...

A: how old are you? (1)

B: I'm 16. (2)

A: 16? (3)

B: I don't know what you are talking about. (4)

A: You don't know what you are saying. (5)

B: I don't know what you are talking about . (6)

A: You don't know what you are saying. (7)

...

Li et al. (2016), "Deep Reinforcement Learning for Dialogue Generation" *EMNLP*



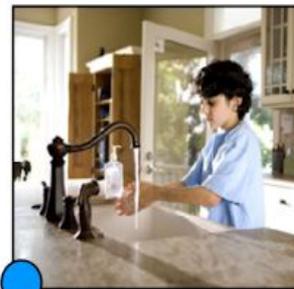
Where are we now?



<https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist>



woman cooking



man fixing faucet

Zhao, J., Wang, T., Yatskar, M., Ordonez, V and Chang, M.-W. (2017) Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraint. EMNLP

Why is NLP Hard?

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

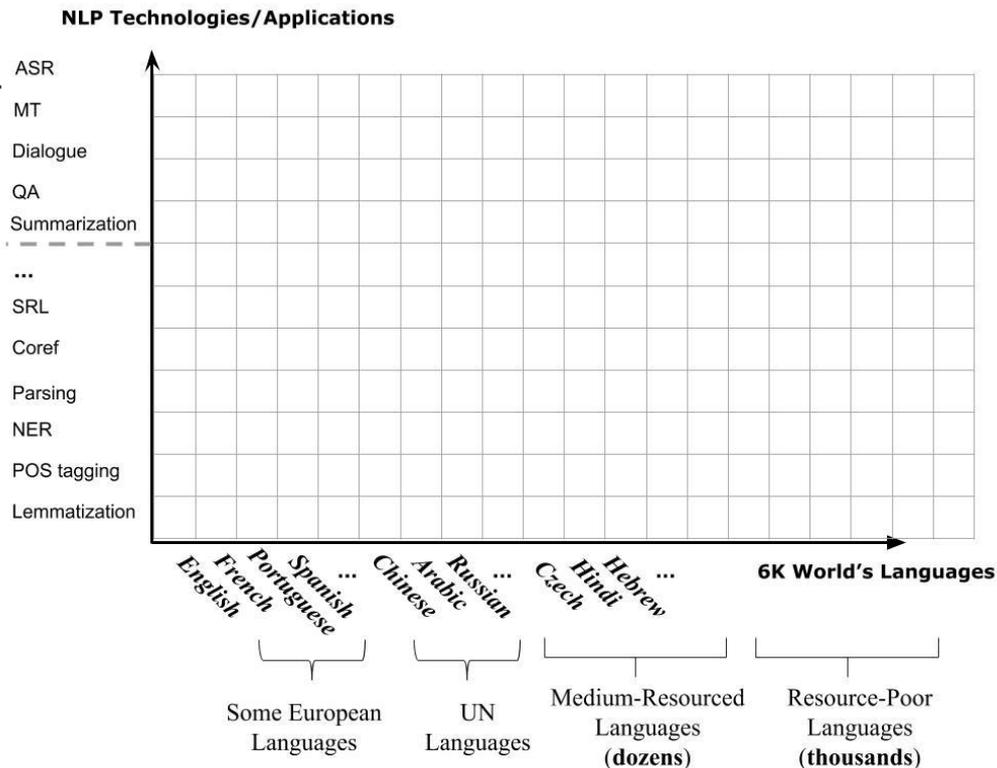


Ambiguity

- **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**



Ambiguity + Scale



Tokenization

这是一个简单的句子

WORDS

This is a simple sentence

זה משפט פשוט



Word Sense Disambiguation

in tea
her daughter

בתה

- most of the vowels unspecified



Tokenization + Disambiguation

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

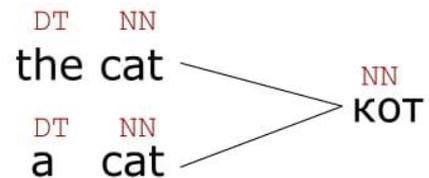
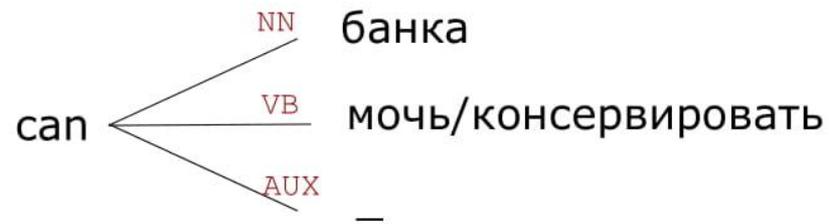
ושבתה

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



Part of Speech Tagging



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



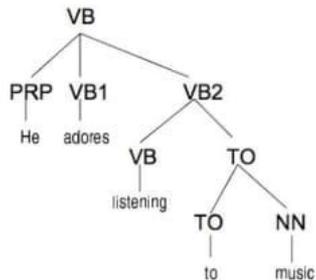
Morphology

unfriend, Obamacare, Manfuckinghattan



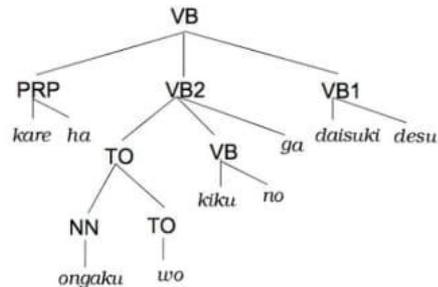
Syntactic Parsing, Word Alignment

SVO

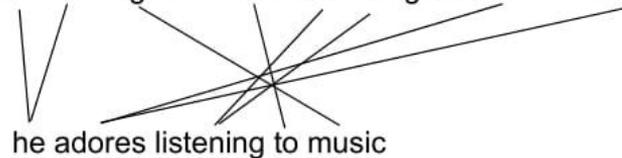


he adores listening to music

SOV



かれは おんがく を きく のが だいすき です
kare ha ongaku wo kiku no ga daisuki desu



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. *it's raining cats and dogs* or *wake up* and metaphors, e.g. *love is a journey* are very different across languages

Semantics

Every fifteen minutes a woman in this country gives birth.

Semantics

Every fifteen minutes a woman in this country gives birth. Our job is to find this woman, and stop her!

– Groucho Marx



Syntax + Semantics

We saw the woman with the telescope wrapped in paper.

- Who has the telescope?
- Who or what is wrapped in paper?
- An event of perception, or an assault?

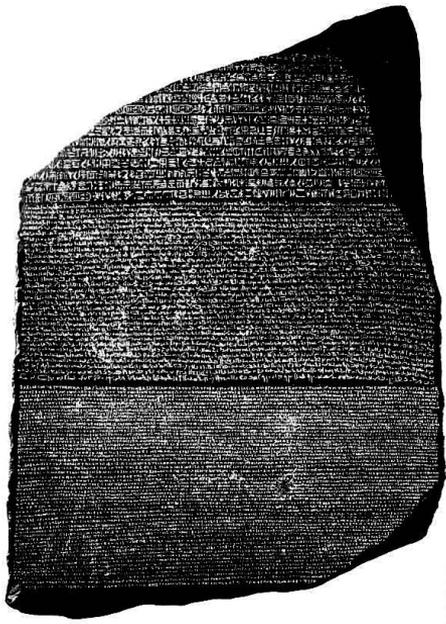
Dealing with Ambiguity

- 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 POS tagging, PCFGs for syntax) and algorithms (Viterbi, probabilistic CKY) return *the best possible analysis*, i.e., the most probable one according to the model.

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



Corpora



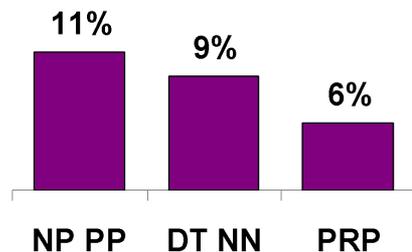
- A corpus is a collection of text
 - Often annotated in some way
 - Sometimes just lots of text
- Examples
 - Penn Treebank: 1M words of parsed WSJ
 - Canadian Hansards: 10M+ words of aligned French / English sentences
 - Yelp reviews
 - The Web: billions of words of who knows what



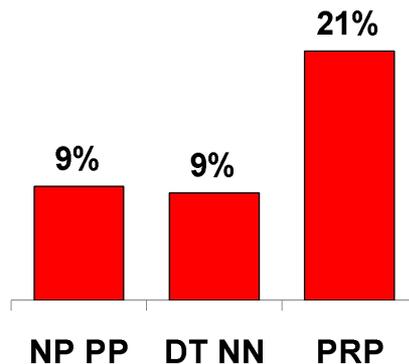
Corpus-Based Methods

- Give us statistical information

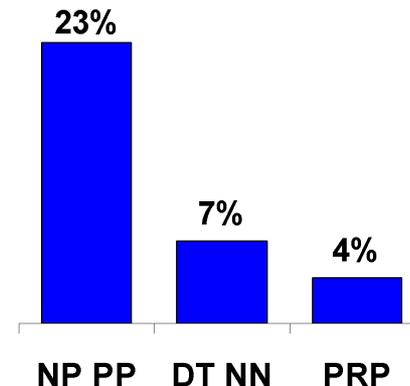
All NPs



NPs under S

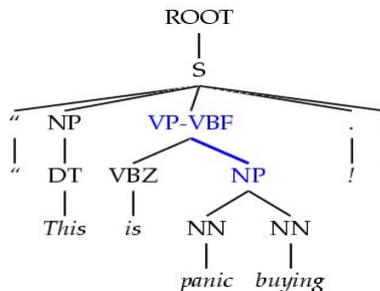
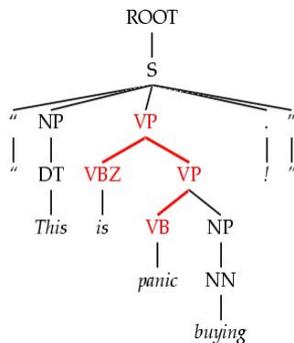


NPs under VP



Corpus-Based Methods

- Let us check our answers



TRAINING

DEV

TEST



Statistical NLP

Like most other parts of AI, NLP is dominated by statistical methods

- Typically more robust than earlier rule-based methods
- Relevant statistics/probabilities are *learned from data*
- Normally requires lots of data about any particular phenomenon



Why is NLP Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled variables
7. Unknown representation



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.

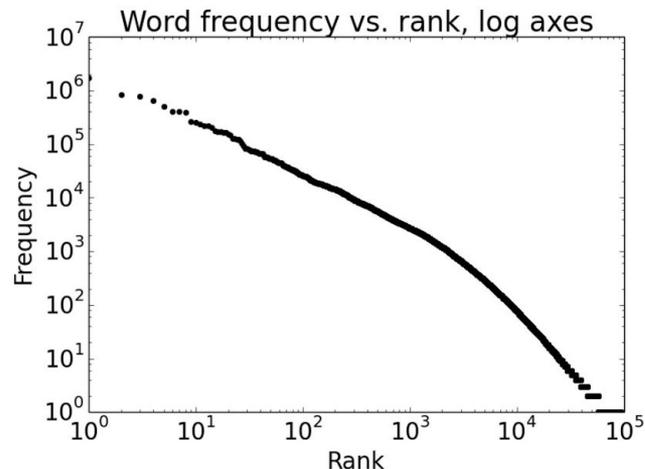
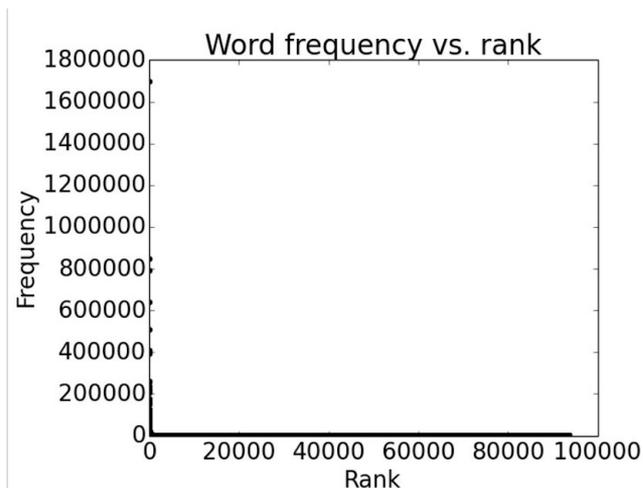
Examples:

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

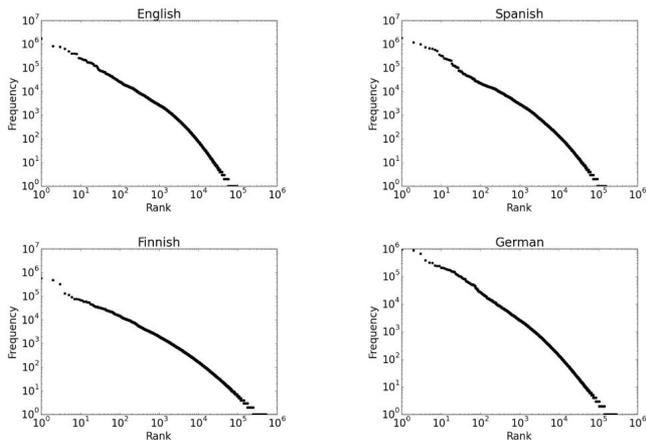


Plotting word frequencies

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



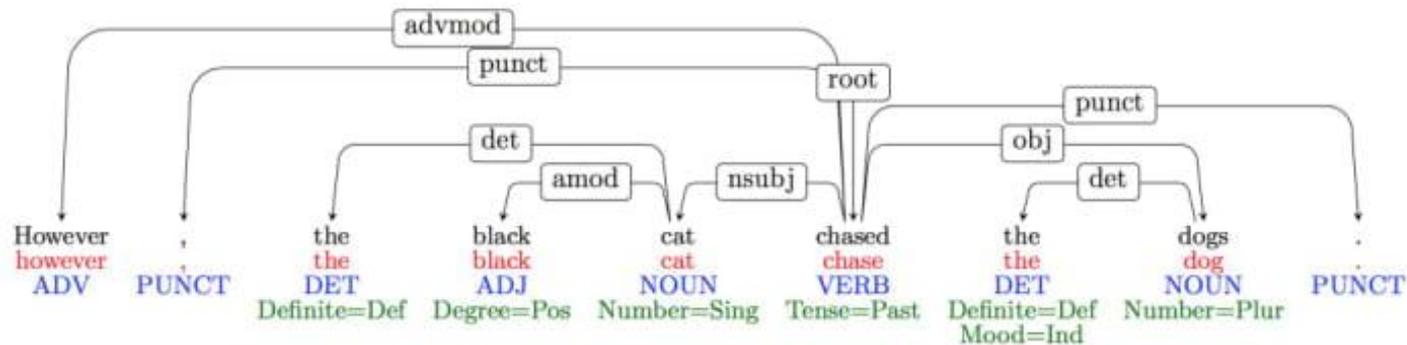
Why is NLP Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled variables
7. Unknown representation



Variation

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

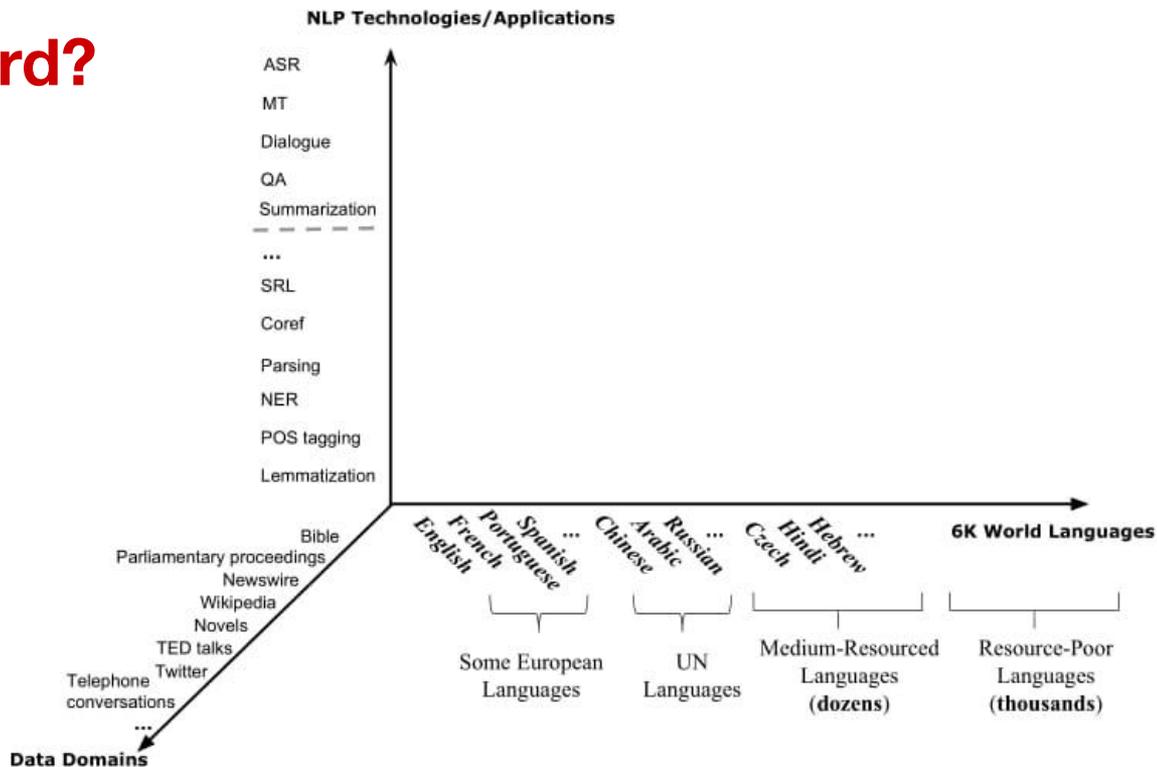


- What will happen if we try to use this tagger/parser for social media??

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



Why is NLP Hard?



Why is NLP Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled variables
7. Unknown representation



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

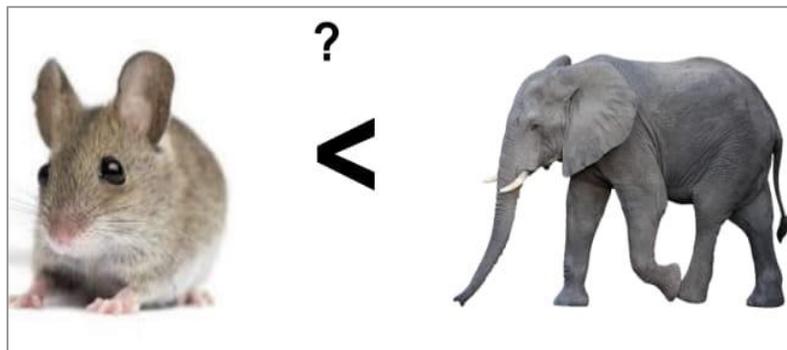
Some kids popped by vs. A few children visited

Is that window still open? vs. Please close the window

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



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?



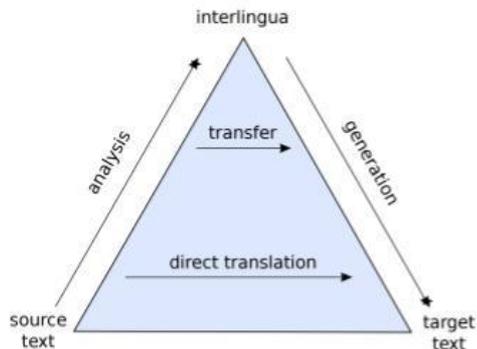
Desiderata for NLP models

- 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

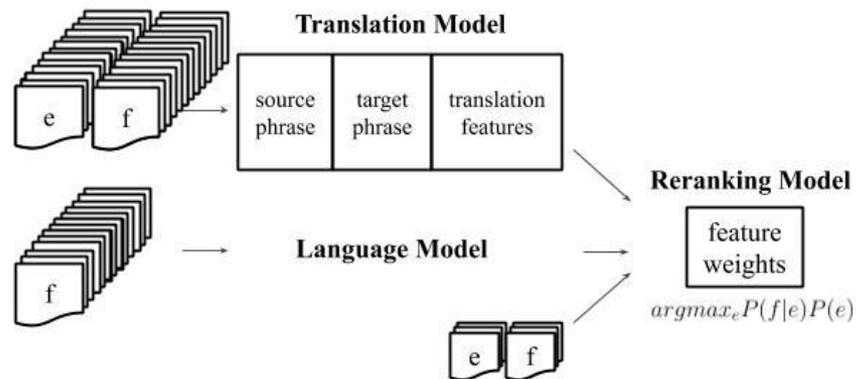
Logic-based/Rule-based NLP



~ 90s

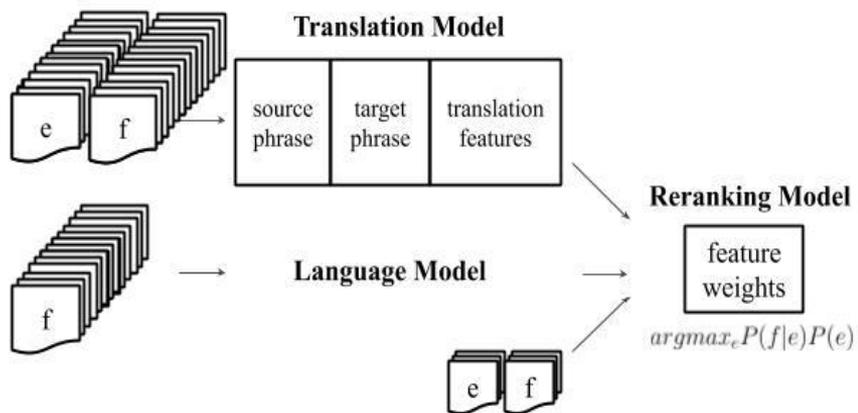


Statistical NLP



Probabilistic and Connectionist NLP

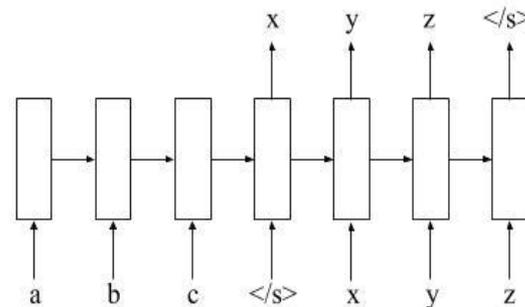
Engineered Features/Representations



~mid 2010s



Learned Features/Representations



NLP $\hat{=}$ 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.



Logistics

What is this Class?

Three aspects to the course:

- **Linguistic Issues**
 - What are the range of language phenomena?
 - What are the knowledge sources that let us disambiguate?
 - What representations are appropriate?
 - How do you know what to model and what not to model?
- **Statistical Modeling Methods**
 - Increasingly complex model structures
 - Learning and parameter estimation
 - Efficient inference: dynamic programming, search, sampling
- **Engineering Methods**
 - Issues of scale, We'll focus on what makes the problems hard, and what works in practice...



What is this Class? Models and Algorithms

- **Models**
 - State machines (finite state automata/transducers)
 - Rule-based systems (regular grammars, CFG, feature-augmented grammars)
 - Logic (first-order logic)
 - Probabilistic models (WFST, language models, HMM, SVM, CRF, ...)
 - Vector-space models (embeddings, seq2seq)
- **Algorithms**
 - State space search (DFS, BFS, A*, dynamic programming---Viterbi, CKY)
 - Supervised learning
 - Unsupervised learning
- **Methodological tools**
 - training/test sets
 - cross-validation



Outline of topics

- Words and Sequences
 - Probabilistic language models
 - Vector semantics and word embeddings
 - Sequence labeling: POS tagging, NER
 - HMMs, Speech recognition
- Structured Classification
- Parsers
- Morphology
- Semantics
- Applications
 - Machine translation, Dialog, Sentiment Analysis



Outline

Aug 27	Course Introduction	Yulia
Aug 29	Language Modeling I	Yulia
Sep 3	Language Modeling II	Yulia
Sep 5	Vector Semantics and Word Embeddings	Yulia
Sep 10	Word Embeddings II	Yulia
Sep 12	POS Tagging, NER	Yulia
Sep 17	HMMs, Speech Recognition I	Yulia
Sep 19	Speech Recognition II	Yulia
Sep 24	Formal Grammars	Bob
Sep 26	Parsing I	Yulia
Oct 1	Parsing II	Yulia
Oct 3	Parsing III	Anjalie
Oct 8	Structured Classification I	Sachin
Oct 10	Structured Classification II	Sachin
Oct 15	Morphology; Features and Unification	Bob
Oct 17	Semantics and Discourse I	Bob
Oct 22	Semantics and Discourse II	Bob
Oct 24	Semantics and Discourse III	Bob
Oct 29	Semantics and Discourse IV	Bob
Oct 31	Machine Translation: Alignment I	Yulia
Nov 5	Machine Translation: Alignment II	Bob
Nov 7	Computational Social Science	Anjalie
Nov 12	Machine Translation: Phrase-Based	Yulia
Nov 14	Machine Translation: Neural	Yulia

Grading

- This is a project based course and grading will be done based on **4 homework assignments** (individual) each contributing to 25% of your final grade.

Projects out of 10 points total:

- 6 Points: Successfully implemented what we asked
- 2 Points: Submitted a reasonable write-up
- 1 Point: Write-up is written clearly
- 1 Point: Substantially exceeded minimum metrics
- Extra Credit: Did non-trivial extension to project



Requirements and Goals

- Class requirements
 - Uses a variety of skills / knowledge:
 - Probability and statistics, graphical models
 - Basic linguistics background
 - Strong coding skills (Java)
 - Most people are probably missing one of the above
 - You will often have to work on your own to fill the gaps
- Class goals
 - Learn the issues and techniques of statistical NLP
 - Build realistic NLP tools
 - Be able to read current research papers in the field



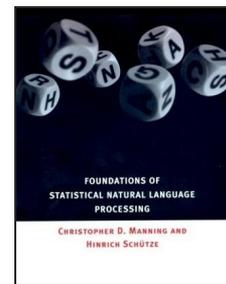
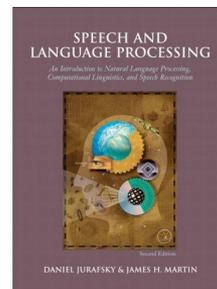
Readings

■ Prerequisites:

- Mastery of basic probability
- Strong skills in Java or equivalent
- Deep interest in language

■ Books:

- Primary text: Jurafsky and Martin, Speech and Language Processing, 2nd and 3rd Edition (not 1st)
<https://web.stanford.edu/~jurafsky/slp3/>
- Also: Manning and Schuetze, Foundations of Statistical NLP
- Also: Eisenstein, Natural Language Processing
<https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf>



Other Announcements

- **Course Contacts:**
 - Webpage: materials and announcements
 - Piazza: discussion forum
 - Canvas: project submissions
 - Homework questions: Recitations, Piazza, TAs' office hours
- **Enrollment: We'll try to take everyone who meets the requirements**
- **Computing Resources**
 - Experiments can take up to hours, even with efficient code
 - Recommendation: start assignments early
- **Questions?**



What's Next?

- Language modeling
 - Start with very simple models of language, work our way up
 - Some statistics concepts that will keep showing up
 - Introduction to machine translation and speech recognition

<http://demo.clab.cs.cmu.edu/11711fa19/>

