Language Technologies Institute

Carnegie Mellon University

Algorithms for NLP

CS 11711, Spring 2020

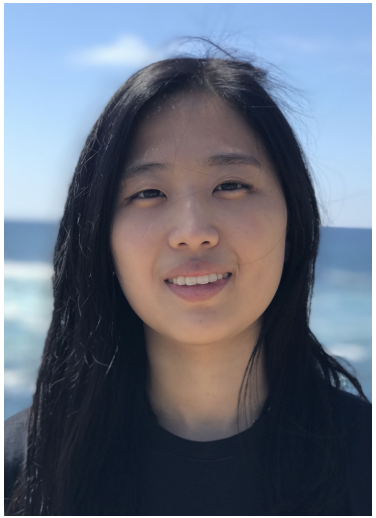
Lecture 1: Introduction

Yulia Tsvetkov

Welcome!

Yulia

David



Chan

Lexi

Course Website

http://demo.clab.cs.cmu.edu/algo4nlp20/

Algorithms for NLP

Spring 2020

T/Th 1:30-3 pm, Wean Hall 2302

Yulia Tsvetkov (office hours: By appointment, GHC 6405), ytsvetko@cs.cmu.edu
David Mortensen (office hours: By appointment, GHC 5407), dmortens@cs.cmu.edu

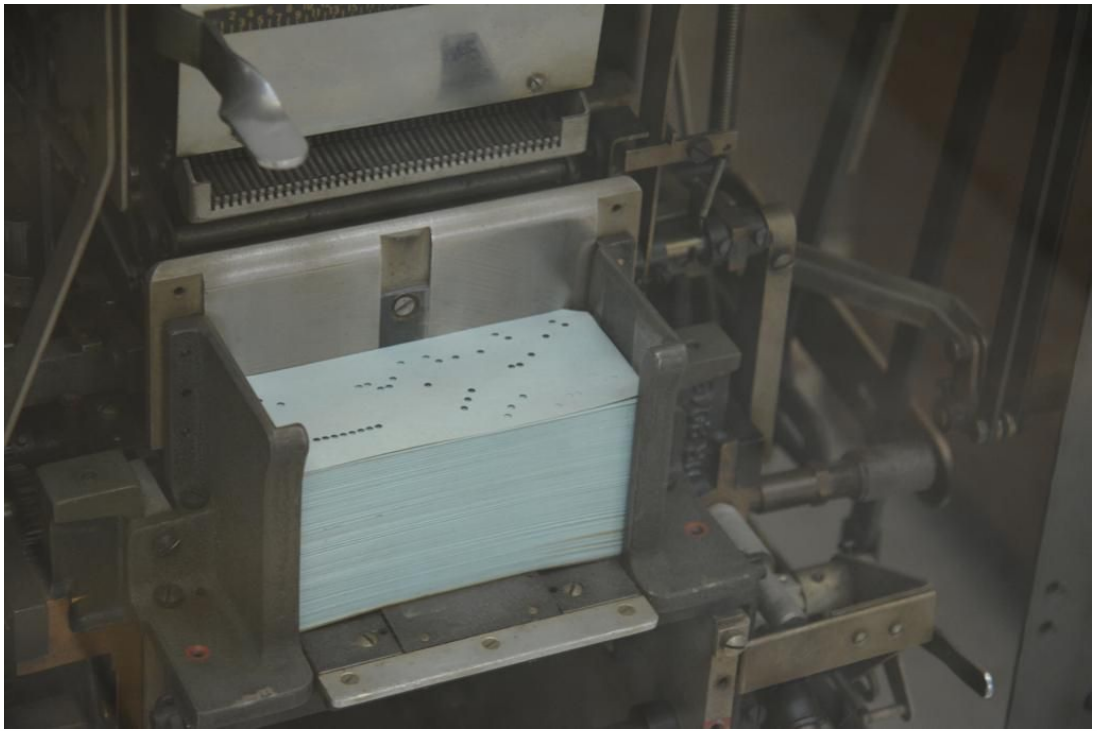
Teaching Assistants:
Chan Young Park (office hours: Thu 4-5pm, TBD), chanyoun@cs.cmu.edu

Forum: Piazza

Summary Announcements Syllabus Readings Grading Policies

Communication with Machines

- ~50s-70s



Communication with Machines

- ~80s

```
File  Edit  Edit_Settings  Menu  Utilities  Compilers  Test  Help
EDIT       BS9U.DEVT3.CLIBPAU(TIMMIES) - 01.31                 Columns 00001 00
Command ===>                                                   Scroll ===> H
****** ***************************** Top of Data ******************************
000001 /* REXX EXEC ***************************************************************
000002 /*
000003 /* TIMMIES FACTOR - COMPOUND INTEREST CALCULATOR
000004 /*
000005 /* AUTHOR: PAUL GAMBLE
000006 /* DATE: OCT 1/2007
000007 /*
000008 /*
000009 /*************************************************************************
000010
000011
000012 say '**********************'
000013 say 'Welcome Coffee drinker.'
000014 say '**********************'
000015 DO WHILE DATATYPE(CoffeeAmt) \= 'NUM'
000016      say ""
000017      say "What is the price of your coffee?",
000018          "(e.g. 1.58 = $1.58)"
000019      parse pull CoffeeAmt
000020 END
000021
000022 DO WHILE DATATYPE(CoffeeWk) \= 'NUM'
000023      say ""
000024      say "How many coffees a week do you have?"
000025      parse pull CoffeeWk
000026 END
000027
000028 DO WHILE DATATYPE(Rate) \= 'NUM'
000029      say ""
000030      say "What annual interest rate would you like to see on that money?",
000031          "(e.g. 8 = 8%)"
000032      parse pull Rate
000033 END
000034 Rate = Rate * 0.01 /* CHG TO DECIMAL NUMBER */
```

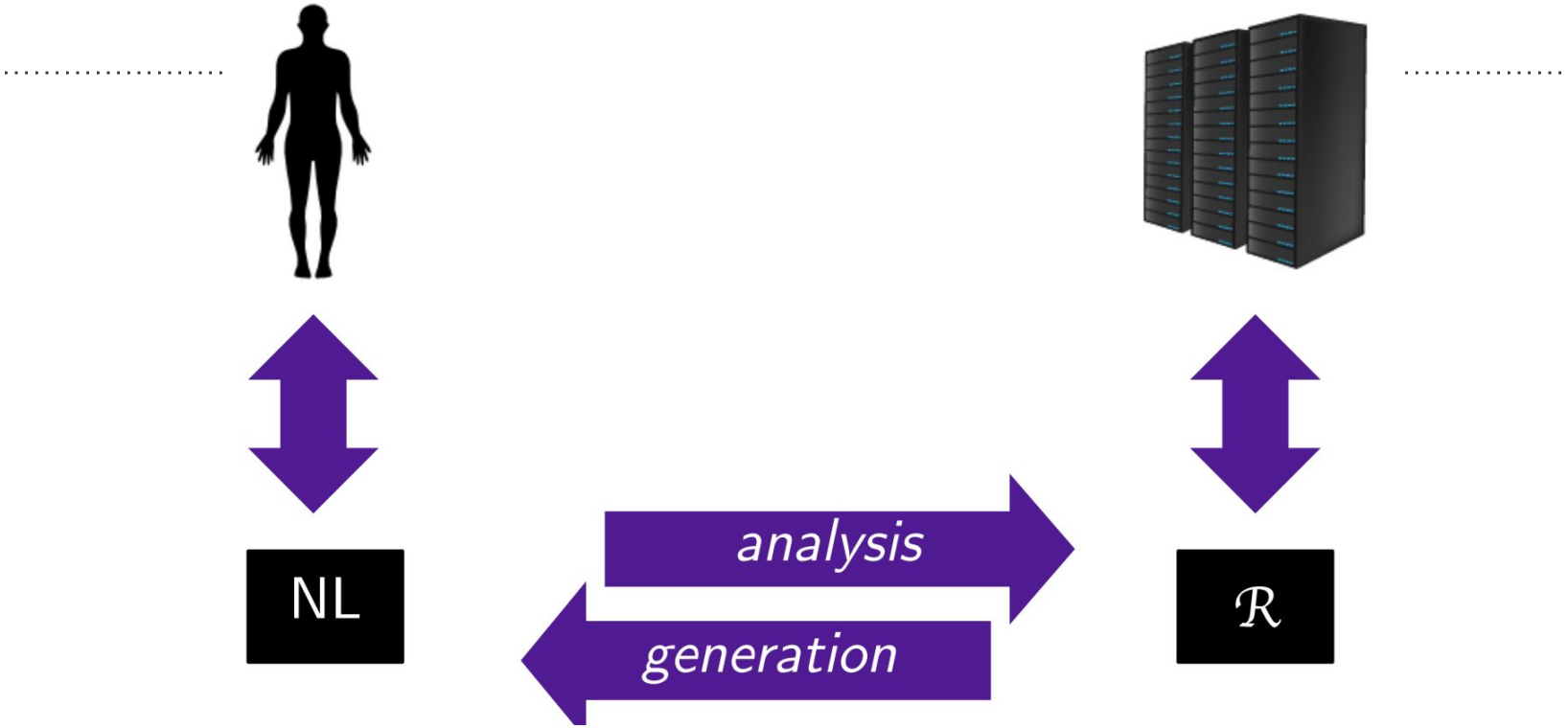
Communication with Machines

- Today

What is NLP?

- NL ∈ {Mandarin, Hindi, Spanish, Arabic, English, … Inuktitut}

- Automation of NLs:
  - analysis (NL $\rightarrow \mathcal{R}$)
  - generation ($\mathcal{R} \rightarrow$ NL)
  - acquisition of $\mathcal{R}$ from knowledge and data



Slide by Noah Smith

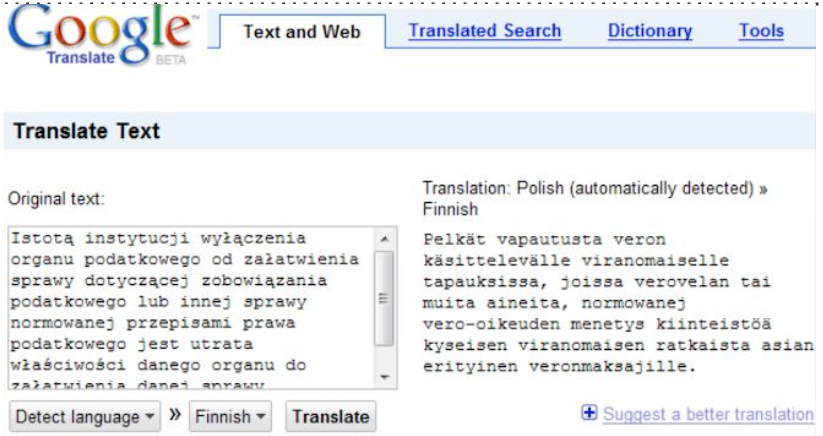
What language technologies are required to write such a program?

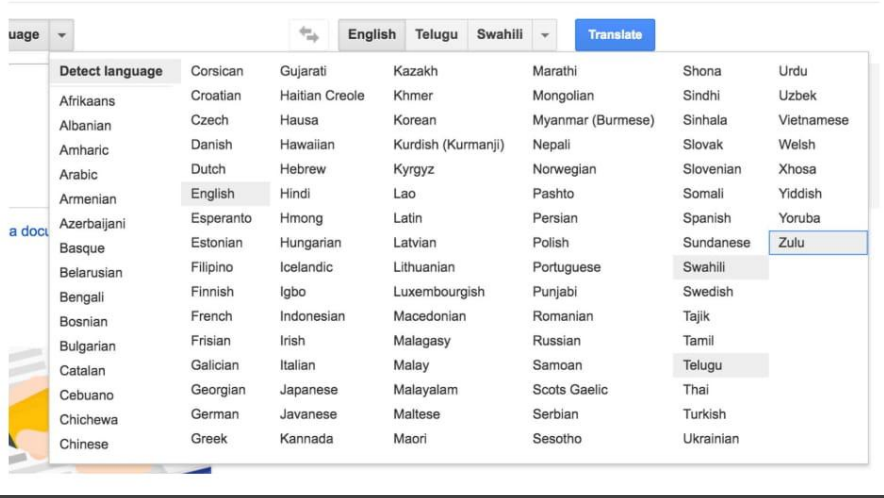
Language Technologies

A conversational agent contains

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

Language Technologies





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

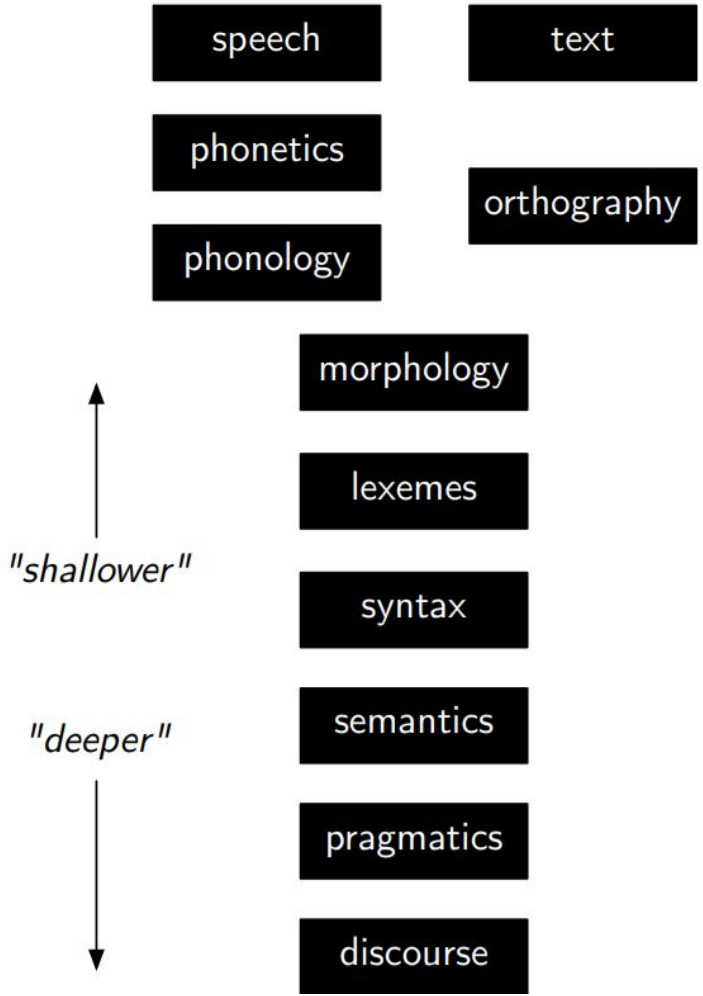
- Applications
  - Machine Translation
  - Information Retrieval
  - Question Answering
  - Dialogue Systems
  - Information Extraction
  - Summarization
  - Sentiment Analysis
  - ...
- Core technologies
  - Language modelling
  - Part-of-speech tagging
  - Syntactic parsing
  - Named-entity recognition
  - 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



Slide by Noah Smith

Phonetics, phonology

- Pronunciation modeling

SOUNDS	Th	i	a	si	e	n

Words

- Language modeling
- Tokenization
- Spelling correction

WORDS This is a simple sentence

Morphology

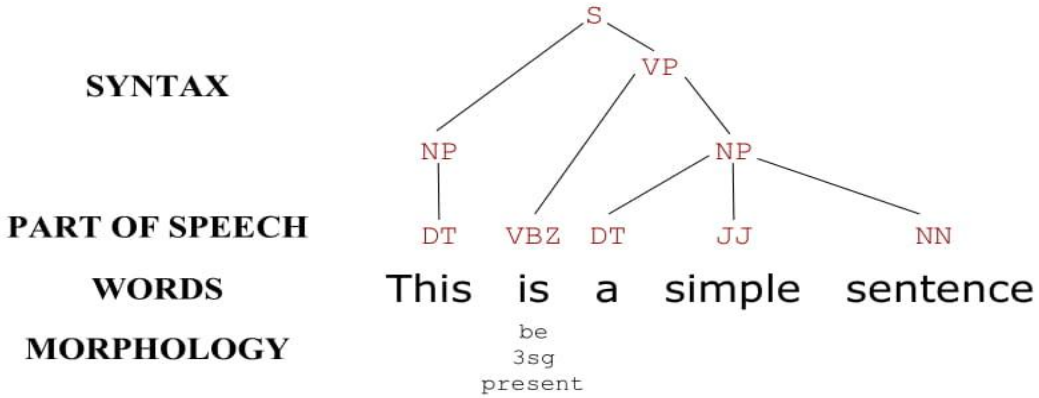
- Morphological analysis
- Tokenization
- Lemmatization

Parts of speech

- Part-of-speech tagging

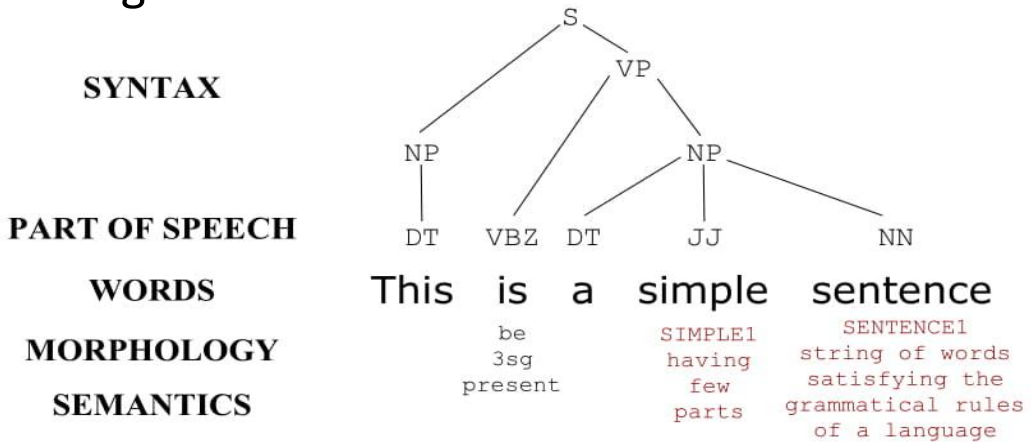
Syntax

- Syntactic parsing



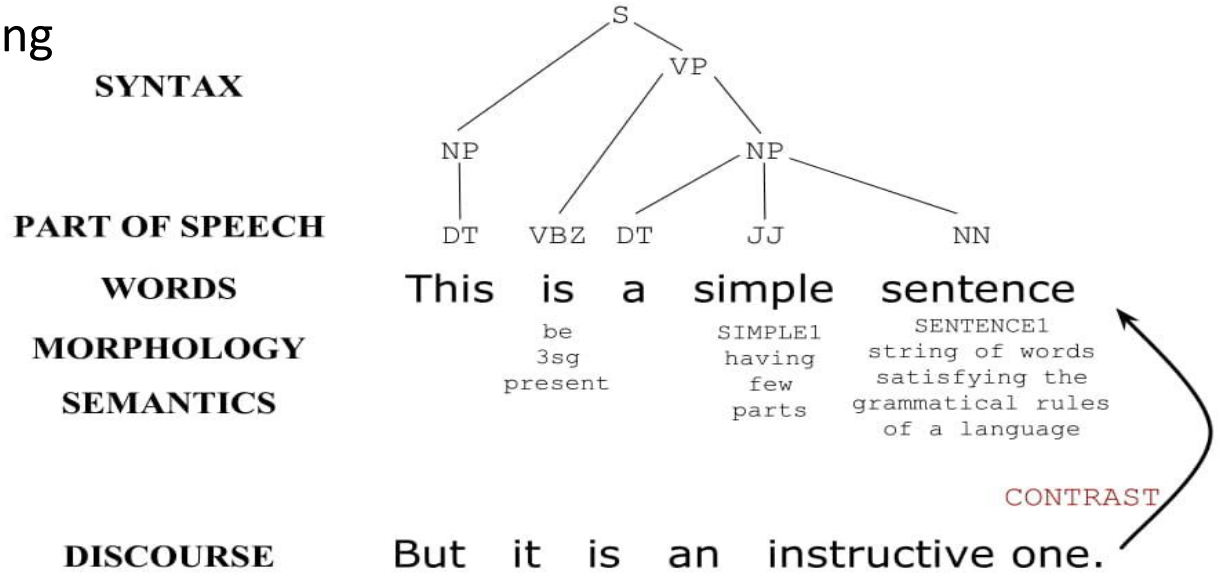
Semantics

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

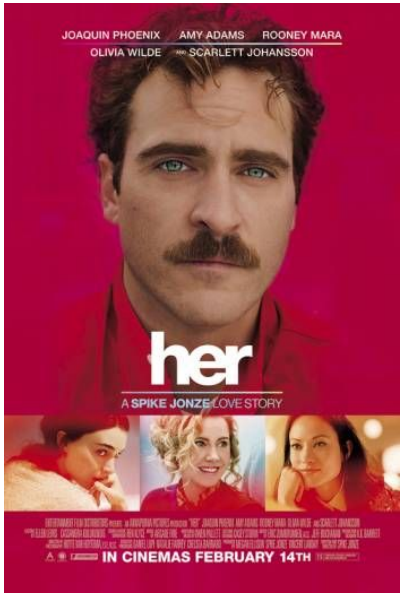


Discourse

- Reference resolution
- Discourse parsing



Where are we now?



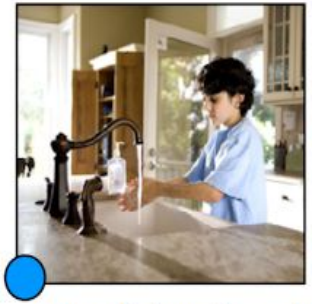
Baseline mutual information model (Li et al. 2015)
A: Where are you going? (1)
B: I'm going to the restroom. (2)
A: See you later. (3)
B: See you later. (4)
A: See you later. (5)
B: See you later. (6)
...
...
A: how old are you? (1)
B: I'm 16. (2)
A: 16? (3)
B: I don't know what you are talking about. (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

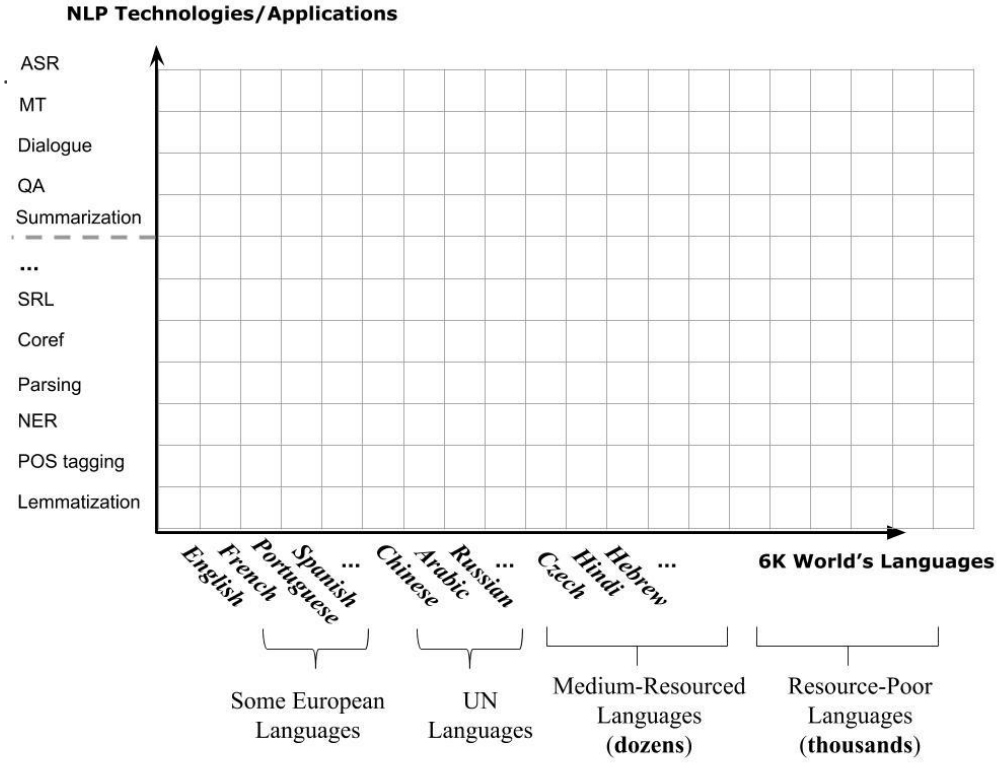
Why is NLP Hard?

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	בתה
her daughter	

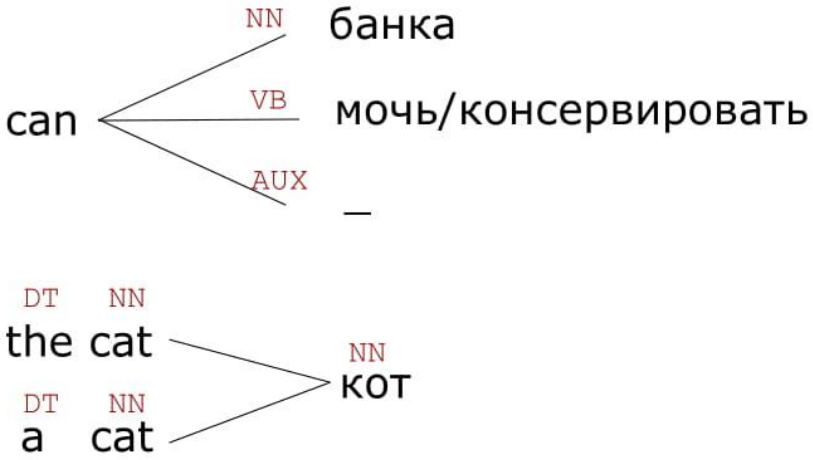
- most of the vowels unspecified

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'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"

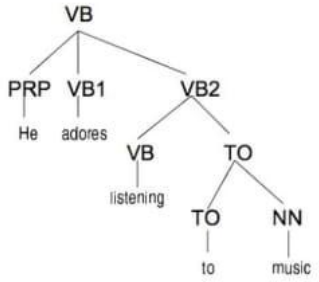
Much'a	to kiss
-na	expresses obligation, lost in translation
-naya	expresses desire
-ka	diminutive
-pu	reflexive (kiss *eachother*)
-sha	progressive (kiss*ing*)
-sqa	declaring something the speaker has not personally witnessed
-ku	3rd person plural (they kiss)
-puni	definitive (really*)
-ña	always
-taq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

Morphology

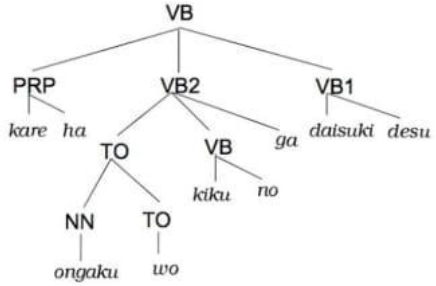
unfriend, Obamacare, Manfuckinghattan

Syntactic Parsing, Word Alignment

SVO



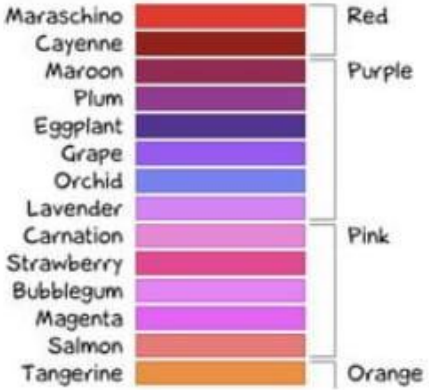
SOV



he adores listening to music

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.

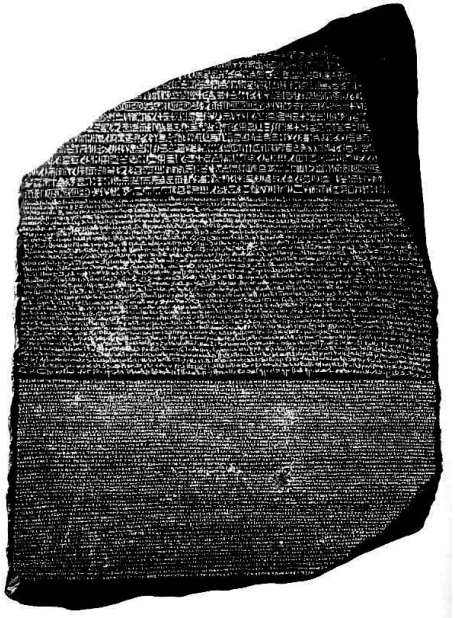
- Who has the telescope?
- Who or what is wrapped in paper?
- An event of perception, or an assault?

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



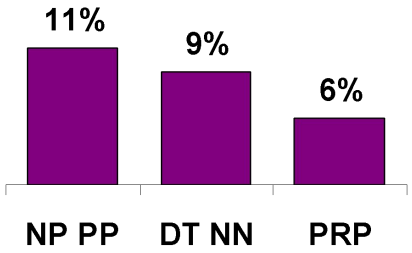
- A corpus is a collection of text
  - Often annotated in some way
  - Sometimes just lots of text

- 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

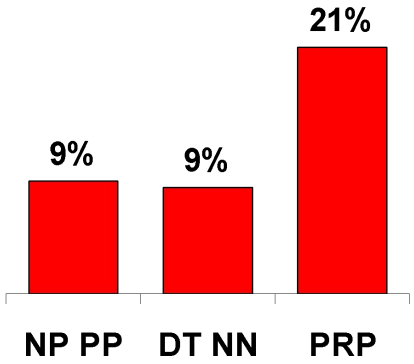
Corpus-Based Methods

- Give us statistical information

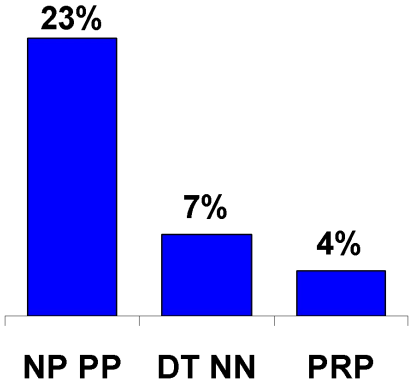
All NPs



NPs under S

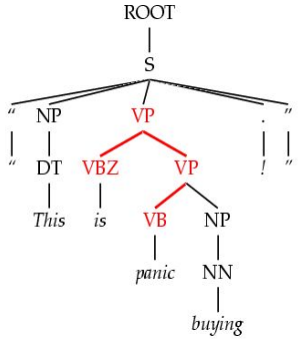


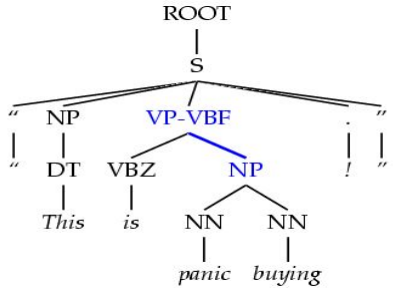
NPs under VP



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

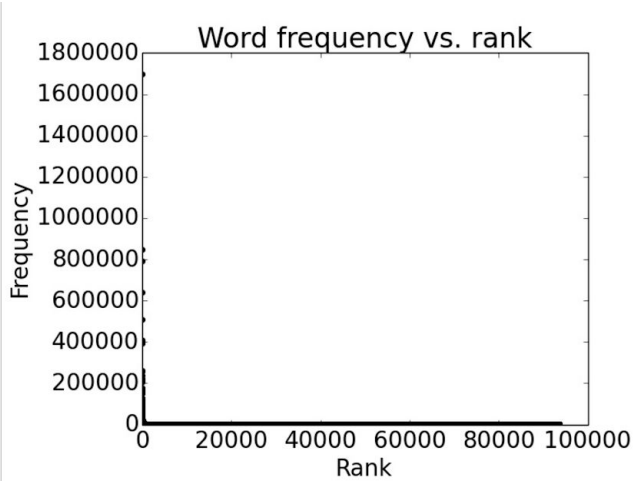
But also, out of 93,638 distinct words (word types), 36,231 occur only once.

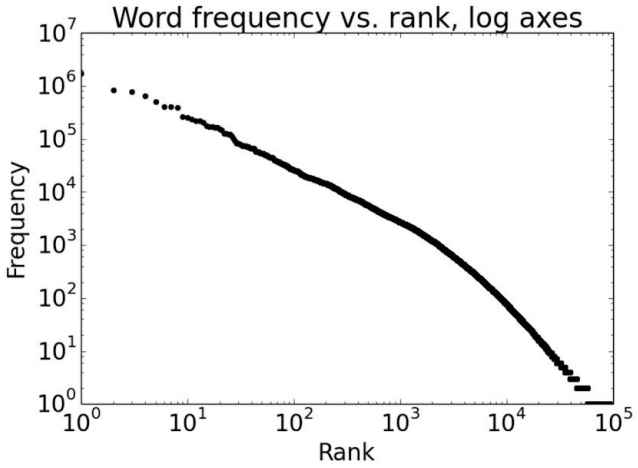
Examples:

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

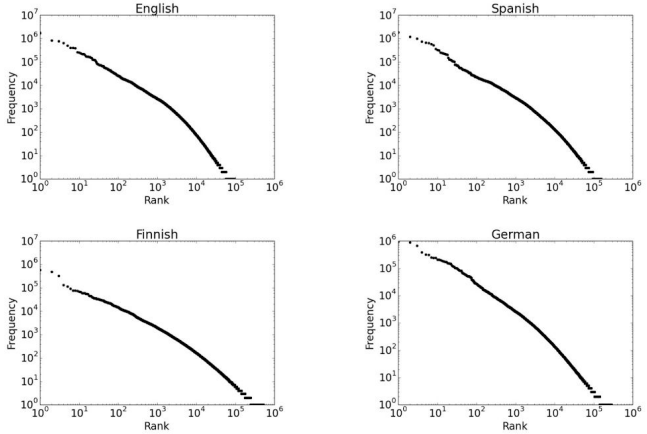
Plotting word frequencies

Order words by frequency. What is the frequency of n_{th} ranked word?





Zipf's Law



Implications

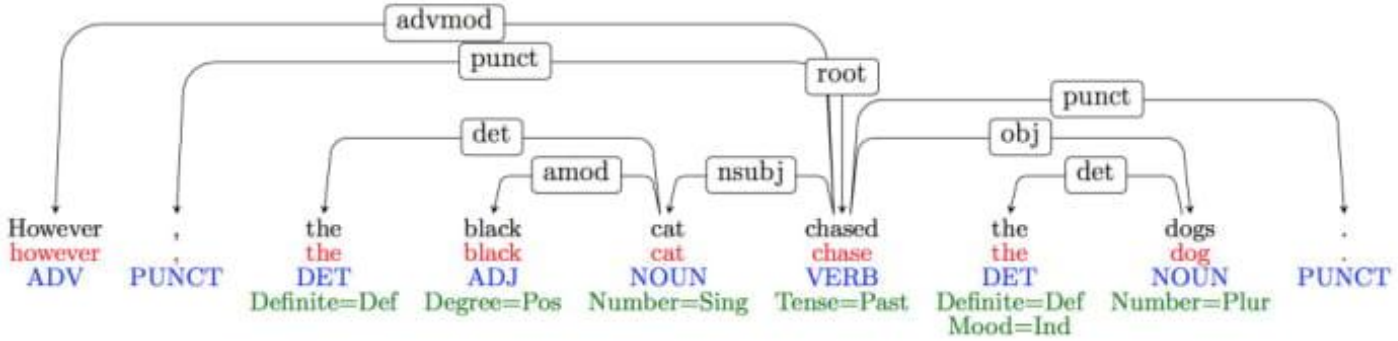
- Regardless of how large our corpus is, there will be a lot of infrequent (and zero-frequency!) words
- This means we need to find clever ways to estimate probabilities for things we have rarely or never seen

Why is NLP Hard?

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

Variation

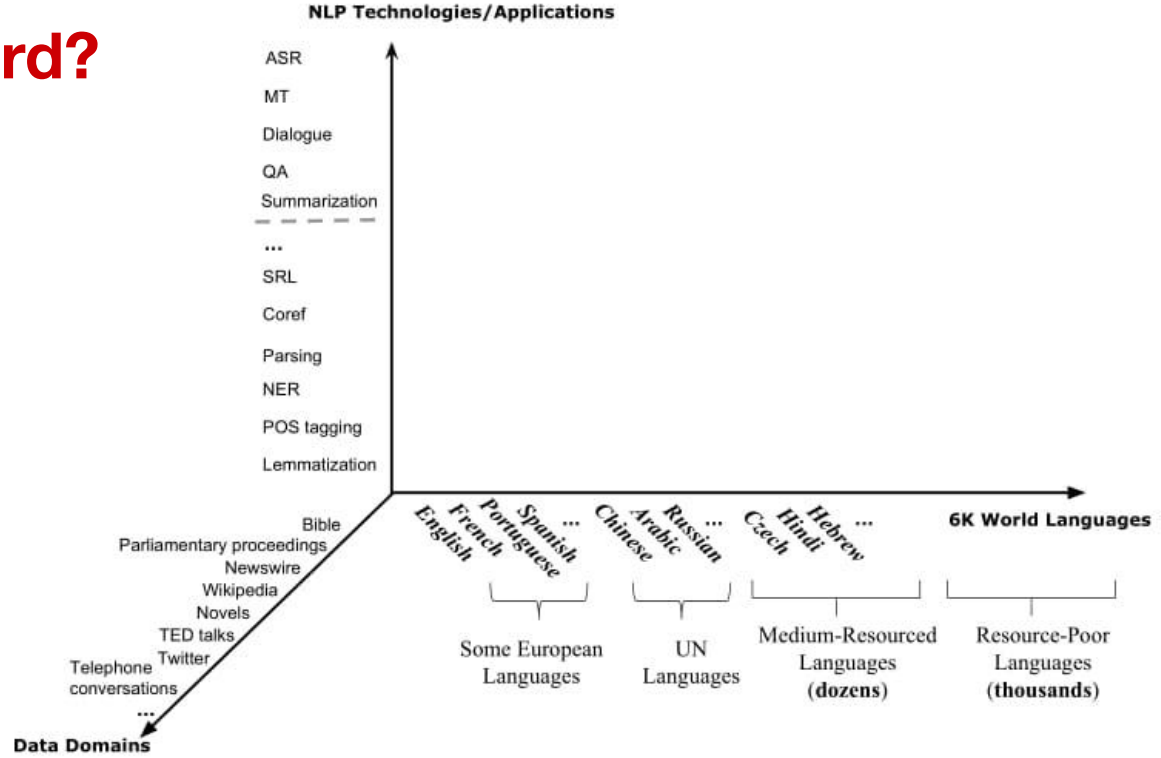
- Suppose we train a part of speech tagger or a parser on the Wall Street Journal



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

@_rkpntrnte 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

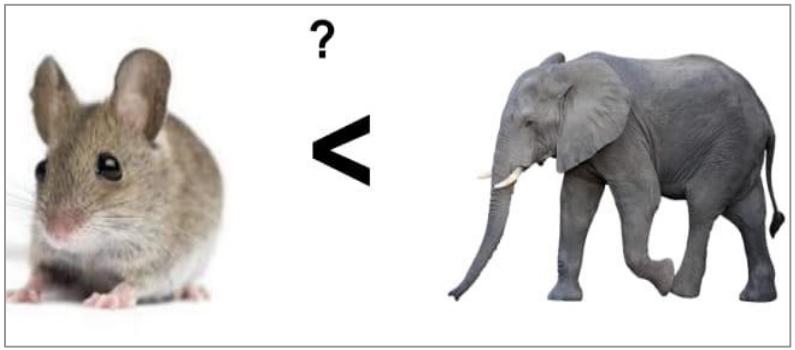
Expressivity

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

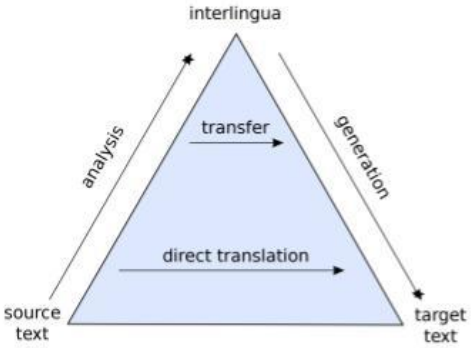
- Very difficult to capture what is $\mathcal{R}$, since we don't even know how to represent the knowledge a human has/needs:
  - What is the "meaning" of a word or sentence?
  - How to model context?
  - Other general knowledge?

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

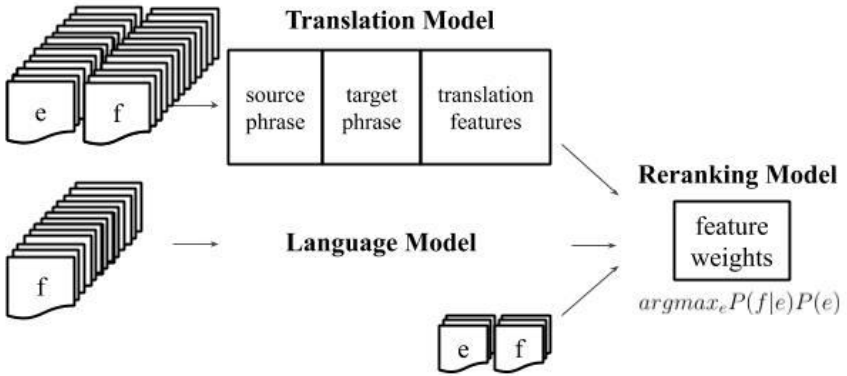
Symbolic and Probabilistic NLP

Logic-based/Rule-based NLP



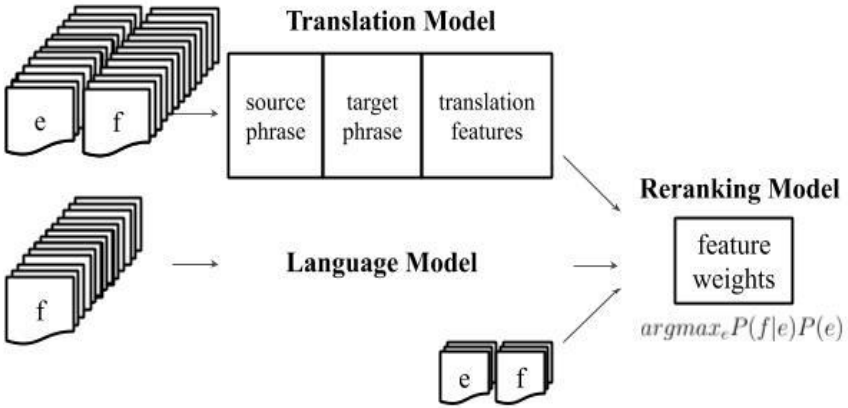
~90s

Statistical NLP



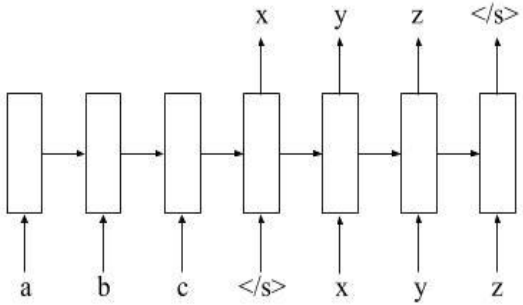
Probabilistic and Connectionist NLP

Engineered Features/Representations



~mid 2010s

Learned Features/Representations

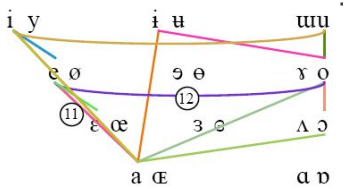


NLP ≟ Machine Learning

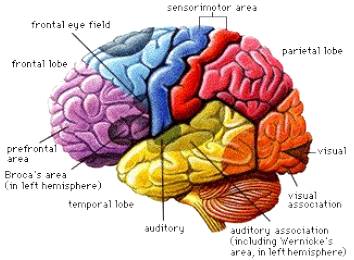
- To be successful, a machine learner needs bias/assumptions; for NLP, that might be linguistic theory/representations.
- Symbolic, probabilistic, and connectionist ML have all seen NLP as a source of inspiring applications.

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

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

Week	Date				
	1/14/20	Intro	Yulia		
1	1/16/20	LM1	Yulia		
	1/21/20	LM2	David		Quiz 1
2	1/23/20	Words	David		
	1/28/20	Word Embeddings I	Yulia		Quiz 2
3	1/30/20	Word Embeddings II	Yulia		
	2/4/20	Lexical Semantics	David		Quiz 3
4	2/6/20	Parts of Speech	David		
	2/11/20	POS Tagging	David		Quiz 4
5	2/13/20	Hmms, POS tagging, NER	Yulia	HW 1 release	
	2/18/20	Formal Grammars	David		Quiz 5
6	2/20/20	Parsing 1	David		
	2/25/20	Parsing 2	Yulia		Quiz 6
7	2/27/20	Parsing 3	Yulia		
	3/3/20	Semantics and Discourse I	David		Quiz 7
8	3/5/20	Semantics and Discourse II	David	HW 1 deadline	
	3/10/20	Spring Break	No class		
9	3/12/20	Spring Break	No class		
	3/17/20	Semantics and Discourse III	David		Quiz 8
10	3/19/20	Semantics and Discourse III	David		
	3/24/20	Pragmatics	David		Quiz 9
11	3/26/20	Sentiment Analysis	Yulia	HW 2 release	
	3/31/20	Neural Network + PyTorch Intro	Jay		
12	4/2/20	Machine Translation	Yulia		Quiz 10
	4/7/20	Machine Translation	Yulia		
13	4/9/20	Spring Carnival	No class		
	4/14/20	Summarization	Chan		
14	4/16/20	Machine Translation	Yulia		
	4/21/20	Low-resource NLP	David	HW 2 deadline	
15	4/23/20				
	4/28/20	Multilingual NLP	David		
16	4/30/20	Ethics	Yulia		

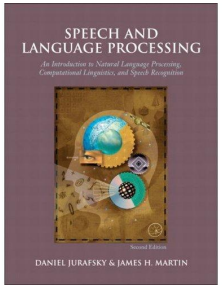
Grading

- Quiz. (50%; 5% each) We'll have a in-class quiz per week, covering the material from the previous week. There will be 10 quizzes in total, and each quiz will take up 5% of the overall grade.

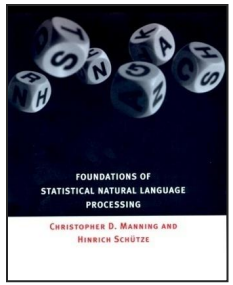
- Project. (50%; 25% each) There will be two Python programming projects; one for POS tagging and one for sentiment analysis. The detailed description on how to submit projects will be given when they are released.

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

- 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/algo4nlp20