COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

Introduction to Natural Language Processing

CMSC 470

Marine Carpuat

Natural Language Processing (NLP)

- The engineering discipline of doing what people do with language, but using computers

Computational Linguistics (CL)

- The science of doing what linguists do with language, but using computers

Speech/Language/Text processing

Human Language Technology

Course staff

Marine Carpuat she/her

Instructor

Assistant Professor

Computer Science

Research: multilingual natural language processing, machine translation

Dennis Asamoah-Owusu

Teaching Assistant

PhD student

Computer Science

Giving computers the ability to process human language is an old idea

Consider the Turing Test, formulated in 1950

- Turing proposed an empirical test/game using language to address a hard question: “How can we determine whether a machine can think?”

- The game
  - 3 participants: two people and a computer
  - the interrogator interacts with 2 other participants via teletype
  - to win the interrogator must determine which of the 2 other participants is the machine

- highlights connection between language, thought and understanding
- but not a rigorous test for intelligence (you can fool some of the people all the time!)

ELIZA program (1966)

- User: You are like my father in some ways.
- ELIZA: WHAT RESEMBLANCE DO YOU SEE

- User: You are not very aggressive but I think you don't want me to notice that
- ELIZA: WHAT MAKES YOU THINK I AM NOT AGGRESSIVE

- User: You don't argue with me
- ELIZA: WHY DO YOU THINK I DON'T ARGUE WITH YOU

- User: You are afraid of me
- ELIZA: DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU

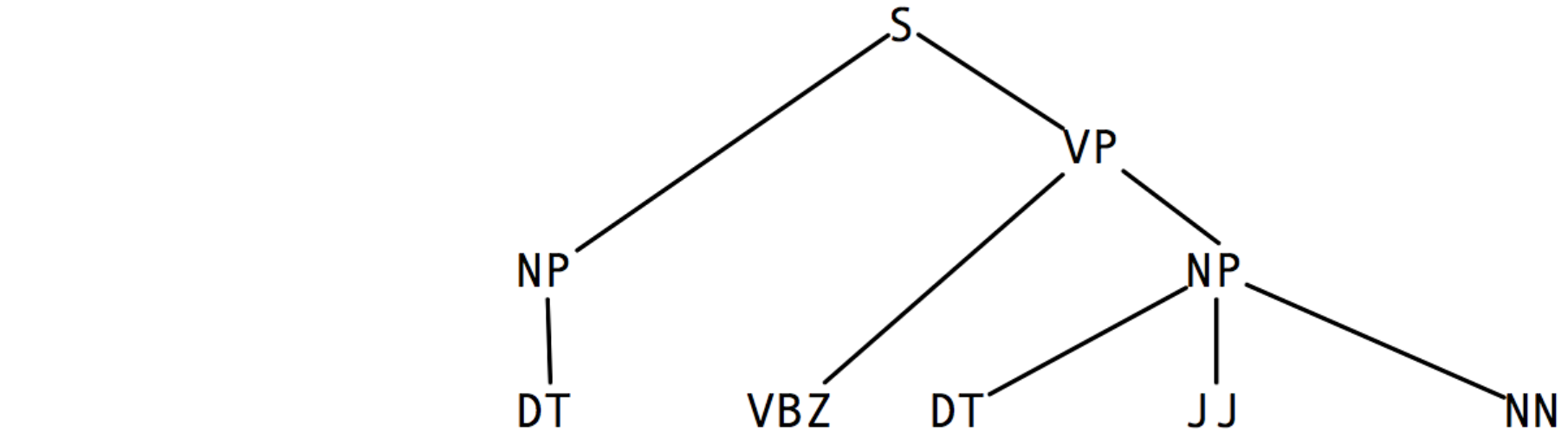
Natural Language Processing Today

What does an NLP system need to “know”?

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

This is a simple sentence

Example from Nathan Schneider



SYNTAX

PART OF SPEECH

This is a simple sentence

WORDS

MORPHOLOGY

be
3sg
present

SIMPLE1
having
few parts

SENTENCE1
string of words
satisfying the
grammatical rules
of a language

SEMANTICS

CONTRAST

But it is an instructive one.

DISCOURSE

Why is NLP hard?

Ambiguity

At the word level

- Part of speech
  - [V Duck]!
  - [N Duck] is delicious for dinner.
- Word sense
  - I went to the bank to deposit my check.
  - I went to the bank to look out at the river

Ambiguity

At the syntactic level

- PP Attachment ambiguity
  - I saw the man on the hill with the telescope
- Structural ambiguity
  - I cooked her duck
  - Visiting relatives can be annoying
  - Time flies like an arrow

Ambiguity

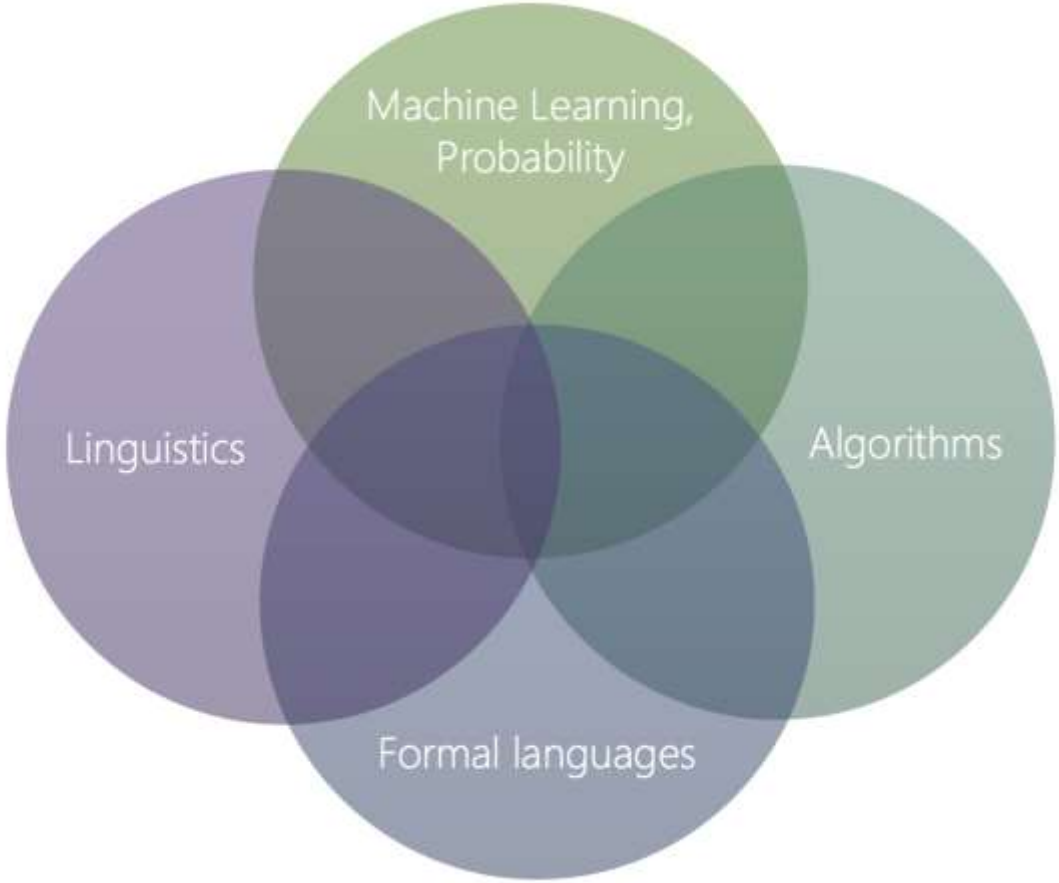
- Quantifier scope
  - Everyone on the island speaks two languages.

- Hard cases require world knowledge, understanding of speaker goals
  - The city council denied the demonstrators the permit because they advocated violence
  - The city council denied the demonstrators the permit because they feared violence

Ambiguity

- NLP challenge: how can we model ambiguity, and choose the correct analysis in context?

- Approach: learn from data



Word counts

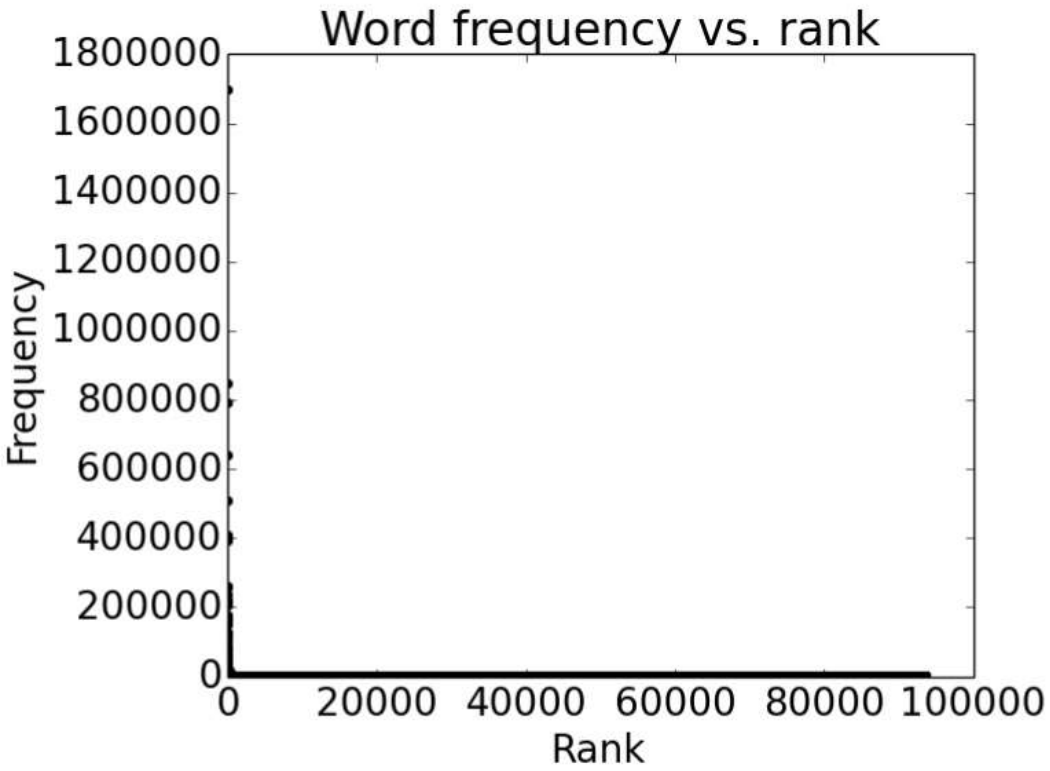
- Most frequent words in the English Europarl **corpus**
- (out of 24M word **tokens**)

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

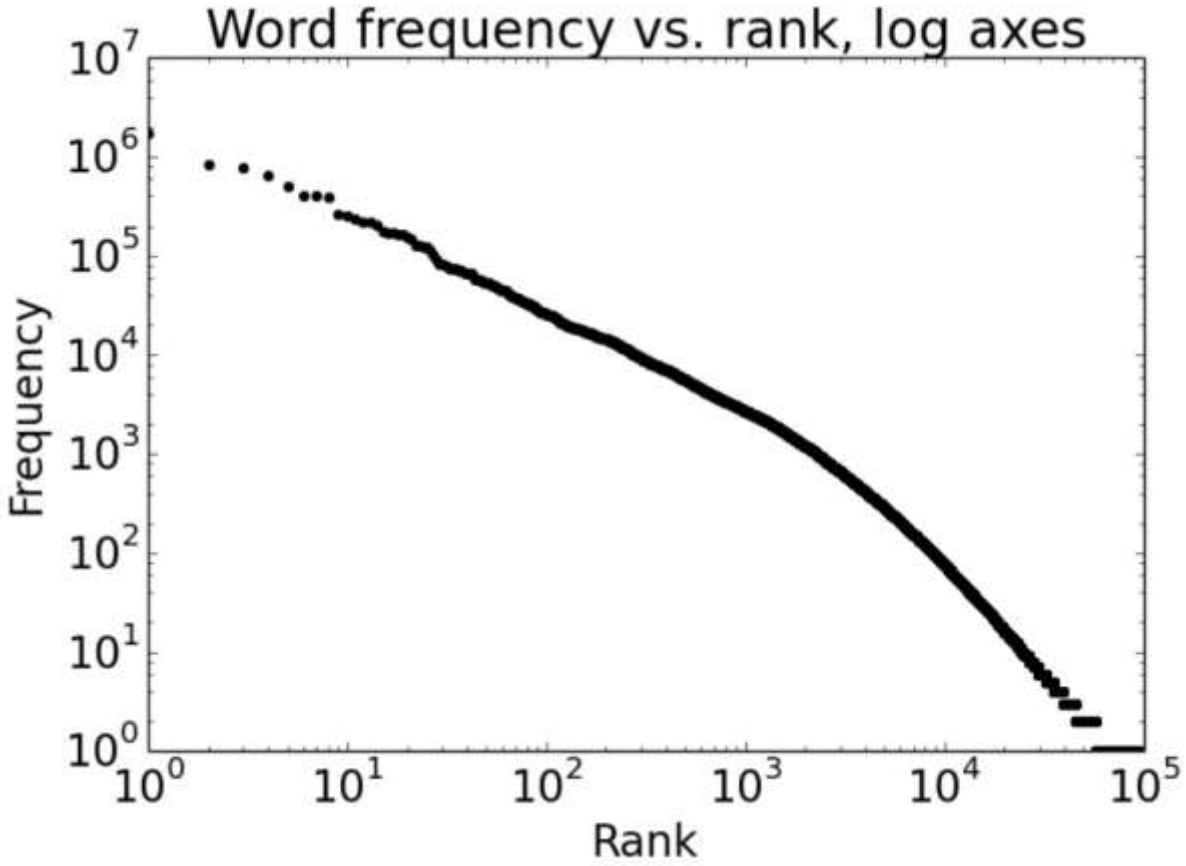
Word counts

- But also, out of the 93,638 distinct words (word **types**), 36,231 occur only once
    - cornflakes, mathematicians, fuzziness, jumbling
    - pseudo-rapporteur, lobby-ridden, perfunctorily,
    - Lycketoft, UNCITRAL, H-0695
    - policyfor, Commissioneris, 145.95, 27a

Plotting word frequencies



Plotting word frequencies (with log-log axes)



Zipf's law

$$f \times r \approx k$$

- $f =$ frequency of a word
- $r =$ rank of a word (if sorted by frequency)
- $k =$ a constant

Zipf's law: implications

- Even in a very large corpus, there will be a lot of infrequent words
- The same holds for many other levels of linguistic structure
- Core NLP challenge: we need to estimate probabilities or to be able to make predictions for things we have rarely or never seen

Variation and Expressivity

- The same meaning can be expressed with different forms

  - I saw the man
  - The man was seen by me

  - She needed to make a quick decision in that situation
  - The scenario required her to make a split-second judgment

6,800 living languages
600 with written tradition
100 spoken by 95% of population

Social Impact

- NLP experiments and applications can have a direct effect on individual users' lives

- Some issues
  - Privacy
  - Exclusion
  - Overgeneralization
  - Dual-use problems

[Hovy & Spruit ACL 2016]

Today's class: what you should know

- Multiple levels of linguistic analysis in NLP
  - Morphology, syntax, semantics, discourse

- Why is NLP hard?
  - Ambiguity
  - Sparse data
    - Zipf's law, corpus, word types and tokens
    - Variation and expressivity
  - Social Impact

This semester

- Words, Context and Meaning
  - Distributional semantics
  - Word sense disambiguation
  - Fundamentals of supervised classification
  - N-gram and neural language models
- Application: Neural Machine Translation
  - Framing and evaluation
  - Neural encoder-decoder models, attention
  - Current research topics
- Linguistic Structure Prediction
  - Sequence labeling tasks
  - Structured prediction and search algorithms
  - Syntax and grammars
  - Parsing

Course Syllabus & Logistics

http://www.cs.umd.edu/class/fall2019/cmsc470/

Exam dates

- Oct 07 3:30pm-4:45pm **Midterm**
- Dec 13 1:30pm-3:30pm **Final**

Before next class

- Read the syllabus
- Check piazza and participate in survey for office hour times
- Get started on homework 1 – due Tuesday Sep 3 by 1:00pm
- Send me a private message on piazza if you are observing religious holidays that overlap with planned exams and assignments