ECE365: Introduction to NLP

Spring 2021

Lecture 1

[Reading J&M 2.2, 2.3, 2.4]

Teaching Staff

Vyom Thakkar (vnt2@illinois.edu)
Office Hours
M, T, Th, Sa, Su: 11am to noon CT
W, Fr: 10am to 11am CT

Suma Bhat (spbhat2@illinois.edu)
Office Hours: By appointment

Logistics

Office hours, reading lists, assignment policies on website

Lab hours + TA office hours will occur virtually via Zoom

Natural Language

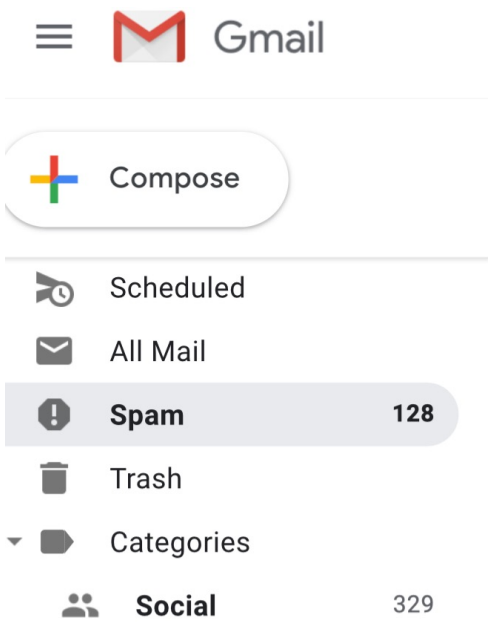
- Basic medium of human communication

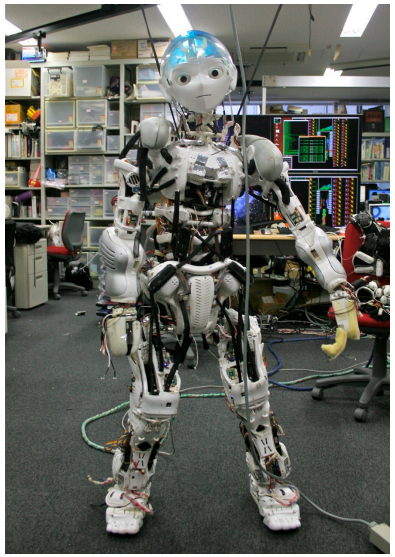
- Language encoded as text
  - Permits broader use of language (in space and time)
  - NL = {Mandarin, Spanish, English,…}

- Mapping between symbols and ideas not always one-to-one
  - Inherent ambiguities (e.g., bank)

Why should machines understand language?

- Human information needs (search, answer questions)





"Pick up the plate with a fork and leave in the sink"

What is Natural Language Processing?

- Making machines do what humans do with language
- In a way they process information
- Design and analysis of algorithms, representations for processing human language

How do machines understand language?

NL → $\mathcal{R}$

$\mathcal{R}$ → Task-specific use

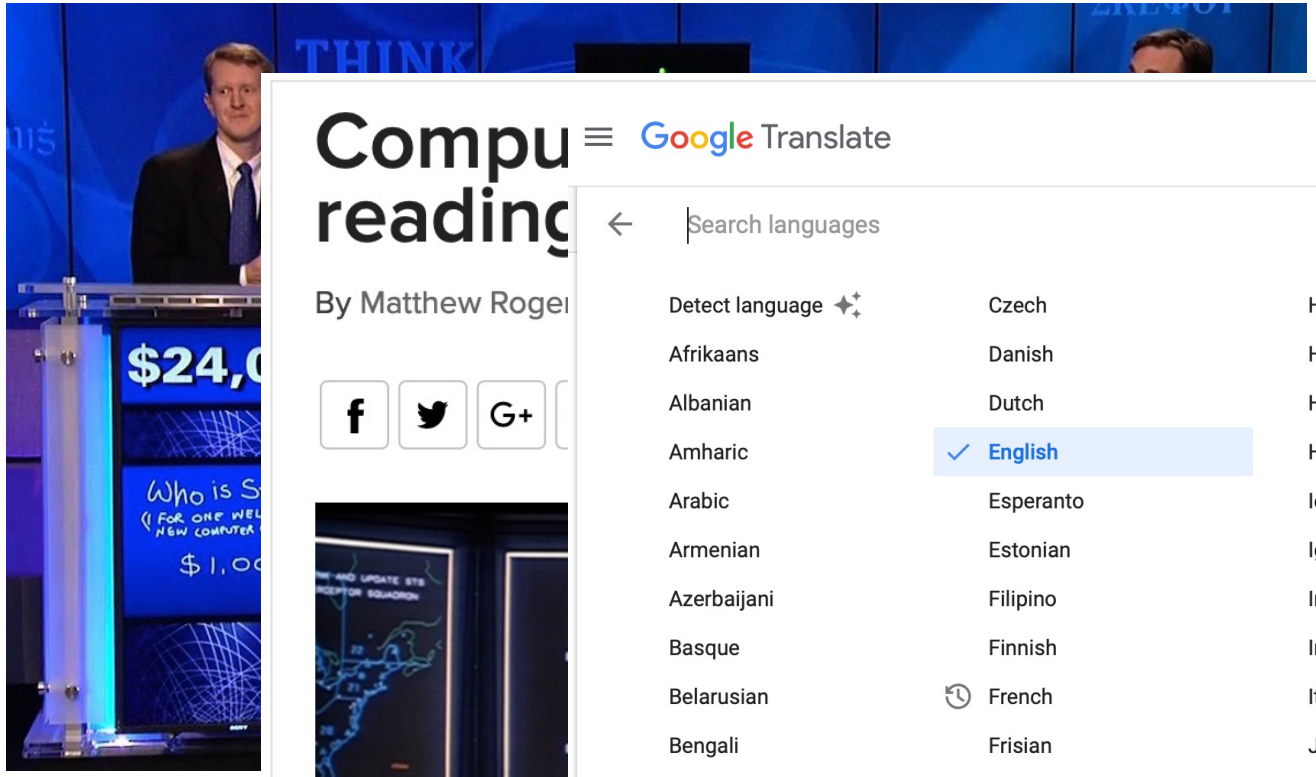
Timing is Right

- Rise of machine learning
- Increased computational capabilities
- Availability of large volumes of textual data
  - Electronic corpora (plural of *corpus* ‘body of text’)

Language Technologies Needed


Hi. I wanted to make an appointment for 8:00 am on Monday. Would the nutritionist be available then?
Yes, the nutritionist is available. May I know the purpose of your visit?
I wanted to go over some of the restrictions of my diet.
Okay, thank you. I'm reserving 8:00 am on Monday for you. May I help you with anything else?

Impressive Feats



Compu
reading

By Matthew Roger

Google Translate

Search languages

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

Language Technologies

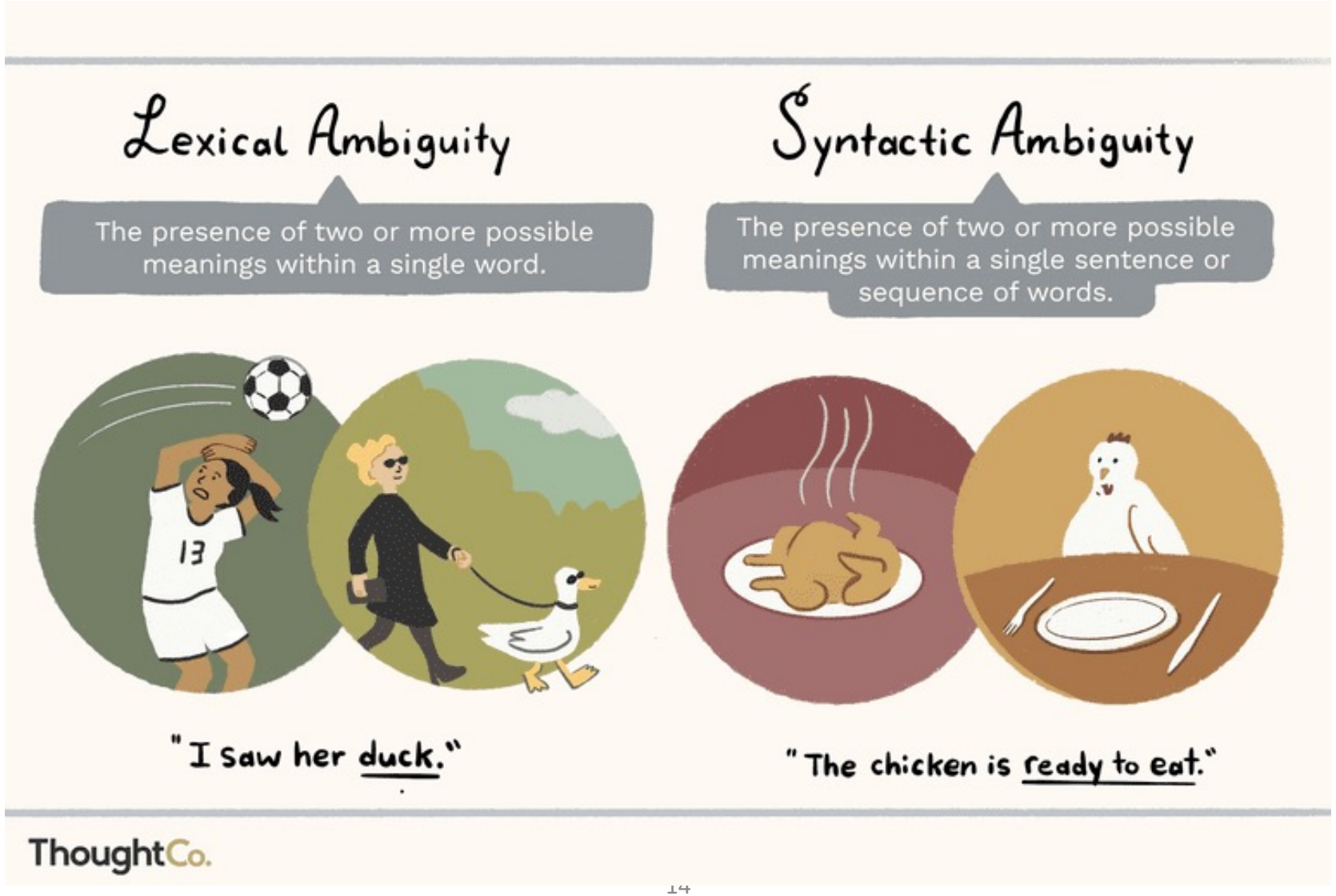
- Applications
  - Machine Translation
  - Information Retrieval
  - Question Answering
  - Dialogue Systems
  - Information Extraction
  - Summarization
  - Sentiment Analysis
  - ...

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

Why is NLP hard?

1. Ambiguities
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unknown representation

Ambiguities



Ambiguities

San Jose cops kill man with knife

Text Paper Translate Listen Close

San Jose cops kill man with knife

Ex-college football player, 23, shot 9 times allegedly charged police at fiancee's home

Who has the knife?

Discourse Ambiguity

Alice invited Bailey for dinner, but she cooked her own food.

- she = Alice or Bailey?

Alice invited Bailey for dinner, but she cooked her own food and brought it with her.

- she = Alice or Bailey?

Alice invited Bailey for dinner, but she cooked her own food and ordered a pizza for her guest.

- she = Alice or Bailey?

Dealing With Ambiguities

- Humans rely on context, common sense

- How can we model ambiguity and choose the correct analysis in context?
  - Probabilistic and non-probabilistic methods

- Where do the models learn from?
  - Large text collections

Learning from Corpora

- Statistical information

- Where do the models learn from?
  - Large text collections

Other Difficulties for NLP

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues

the New York-New Haven Railroad

the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.
Mary and Sue are mothers.

tricky entity names

Where is *A Bug's Life* playing ...
Let It Be was recorded ...
... a mutation on the *for* gene ...

Course Goals

- Understand fundamentals of some sub-fields within NLP (Text classification, Part of speech tagging, Language modeling, vector models of meaning)
- Understand key theories and algorithms for statistical NLP
- Hands on experience building statistical models for language processing
- Same requirements as previous modules in the course

What this course is not

- We will not focus on deep-learning methods
- Will not work with plug-and-play NLP models

Text as Signal

- Text is discrete
  - Meaning created from groups of symbolic units

- What are units of text?

Words as Text Units

- I do uh main- mainly data analysis.
  - Filled pause, fragments

- Seuss's cat is different from other cats.

  Lemma: same stem, part of speech, ~meaning

  cat and cats = same lemma (cat)

  Word form: full inflected surface form

  cat and cats = different forms

Words as Text Units

- They enjoyed the food in New York city but not the stay.
  How many words?
  - Word type: an element of the vocabulary
  - Word token: an instance of the type in the text

	Tokens = N	Types = |V|
Switchboard phone conversations	2.4 million	20 thousand
Shakespeare	884,000	31 thousand
Google N-grams	1 trillion	13 million

Issues in Tokenization

- Tokenization is splitting running text into pieces
  - Friends, Romans, Countrymen, lend me your ears;

Friends
Romans
Countrymen
lend
me
your
ears

Issues with Tokenization

- Finland’s capital – Finlands Finland Finland’s
- What’re, I’m, isn’t – What are, I am, is not
- Lowercase – lowercase or lower case
- San Francisco – two tokens or one?
- m.p.h - ??

Tokenization in Other Languages

- French
  - L'ensemble = Le ensemble, L ensemble, L' ensemble
  - Be able to match with *un ensemble*
- German

  Lebensversicherungsgesellschaftsangestellter
  "Life insurance company employee"

  Need compound splitting

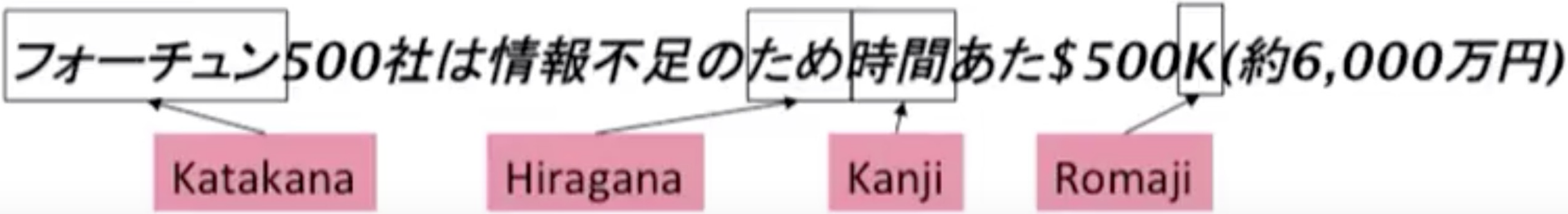
Tokenization in Other Languages

- Chinese

- 莎拉波娃现在居住在美国东南部的佛罗里达。
- 莎拉波娃 现在 居住 在 美国 东南部 的 佛罗里达
- Sharapova now lives in US southeastern Florida

Tokenization in Other Languages

- Japanese



Normalization

- Define equivalence classes of terms
  - Index and query terms need to have same form
  - U.S.A = USA

Case Folding

- Application dependent
  - IR - Reduce to lower case, except when middle of sentence
    - Fed vs fed

  - Sentiment analysis, machine translation – retain case
    - US vs us

Lemmatization

- Find the dictionary headword form
  - Reduce inflections to the base form
  - Am, are, is → be
  - Car, cars, car's, cars' → car

the girl's cars are of different colors →

Stemming

- Reduce terms to their stems

  - word =
  stem (core meaning bearing unit) + affix (attached to stem for grammatical function)

  - Language dependent
  - chopping affix
    - Automate, automatic, automation → automat

    - Porter's Stemming algorithm for English

Summary

- Tokenization
- Normalization
- Stemming
- Lemmatization