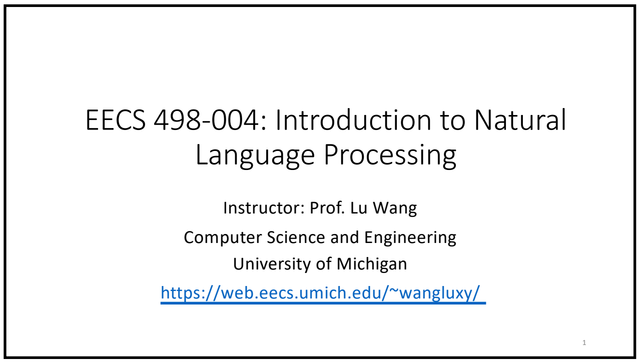


EECS 498-004: Introduction to Natural Language Processing

Instructor: Prof. Lu Wang

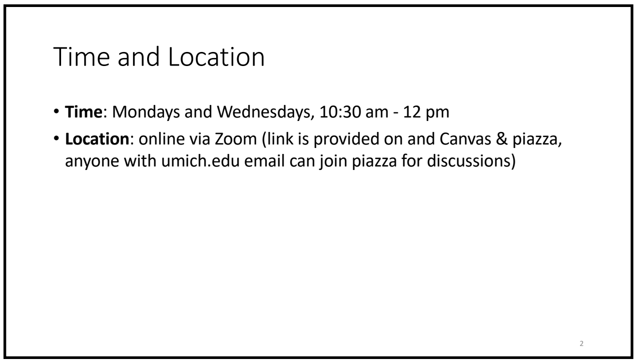
Computer Science and Engineering

University of Michigan

https://web.eecs.umich.edu/~wangluxy/

1

1

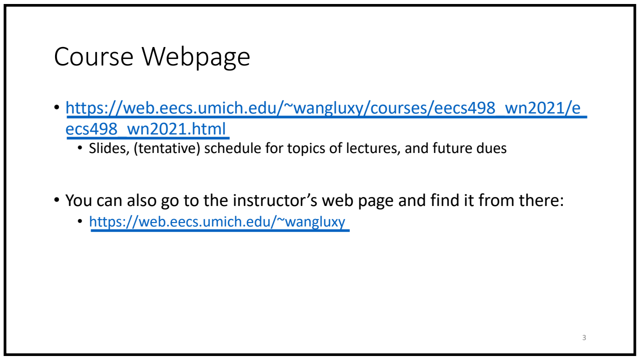


Time and Location

- **Time**: Mondays and Wednesdays, 10:30 am - 12 pm
- **Location**: online via Zoom (link is provided on and Canvas & piazza, anyone with umich.edu email can join piazza for discussions)

2

2

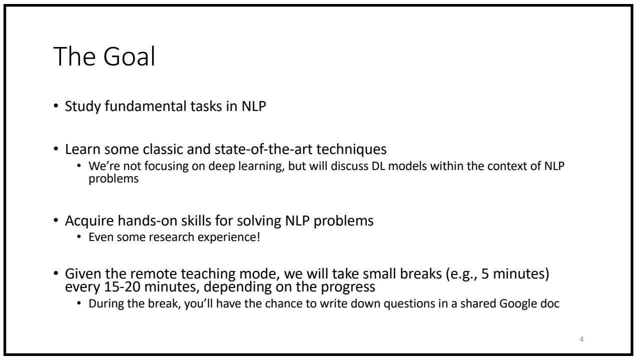


Course Webpage

- https://web.eecs.umich.edu/~wangluxy/courses/eecs498_wn2021/eecs498_wn2021.html
  - Slides, (tentative) schedule for topics of lectures, and future dues
- You can also go to the instructor’s web page and find it from there:
  - https://web.eecs.umich.edu/~wangluxy

3

3

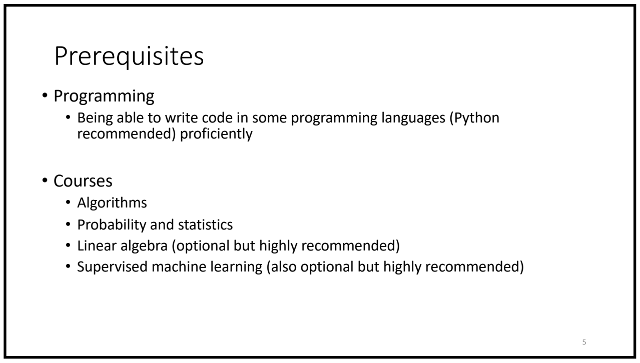


The Goal

- Study fundamental tasks in NLP
- Learn some classic and state-of-the-art techniques
  - We’re not focusing on deep learning, but will discuss DL models within the context of NLP problems
- Acquire hands-on skills for solving NLP problems
  - Even some research experience!
- Given the remote teaching mode, we will take small breaks (e.g., 5 minutes) every 15-20 minutes, depending on the progress
  - During the break, you’ll have the chance to write down questions in a shared Google doc

4

4

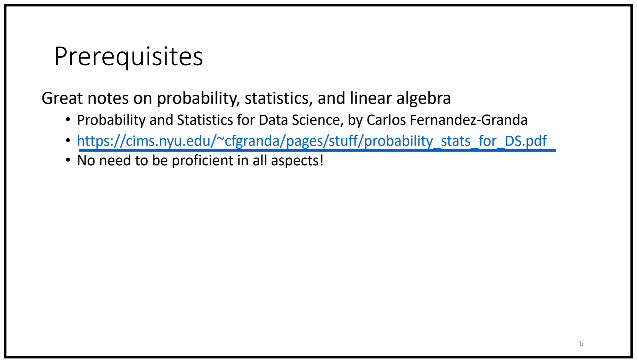


Prerequisites

- Programming
  - Being able to write code in some programming languages (Python recommended) proficiently
- Courses
  - Algorithms
  - Probability and statistics
  - Linear algebra (optional but highly recommended)
  - Supervised machine learning (also optional but highly recommended)

5

5



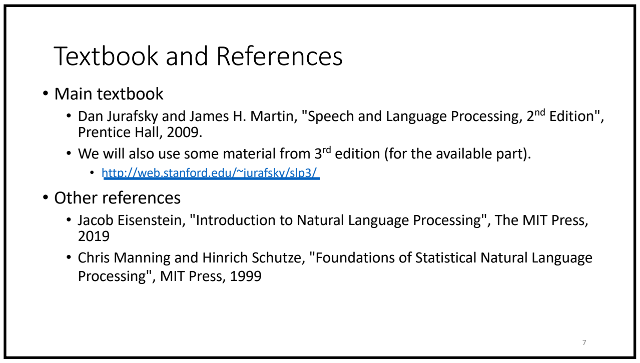
Prerequisites

Great notes on probability, statistics, and linear algebra

- Probability and Statistics for Data Science, by Carlos Fernandez-Granda
- https://cims.nyu.edu/~cfgranda/pages/stuff/probability_stats_for_DS.pdf
- No need to be proficient in all aspects!

6

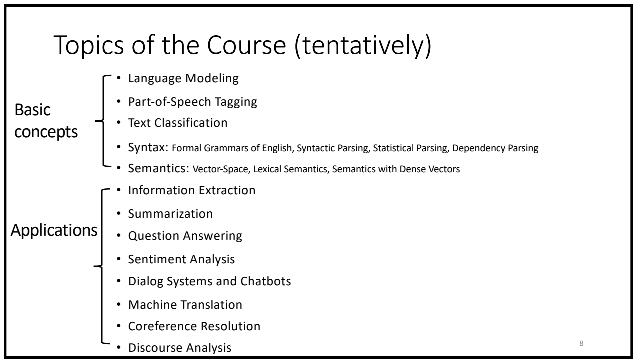
6



Textbook and References

- Main textbook
  - Dan Jurafsky and James H. Martin, "Speech and Language Processing, 2nd Edition", Prentice Hall, 2009.
  - We will also use some material from 3rd edition (for the available part).
    - http://web.stanford.edu/~jurafsky/slp3/
- Other references
  - Jacob Eisenstein, "Introduction to Natural Language Processing", The MIT Press, 2019
  - Chris Manning and Hinrich Schutze, "Foundations of Statistical Natural Language Processing", MIT Press, 1999

7

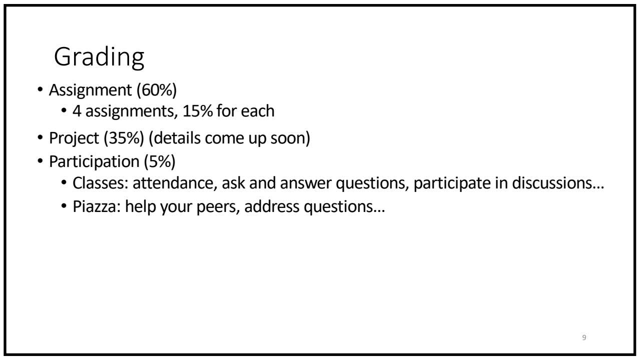


Topics of the Course (tentatively)

Basic concepts
- Language Modeling
- Part-of-Speech Tagging
- Text Classification
- Syntax: Formal Grammars of English, Syntactic Parsing, Statistical Parsing, Dependency Parsing
- Semantics: Vector-Space, Lexical Semantics, Semantics with Dense Vectors

Applications
- Information Extraction
- Summarization
- Question Answering
- Sentiment Analysis
- Dialog Systems and Chatbots
- Machine Translation
- Coreference Resolution
- Discourse Analysis

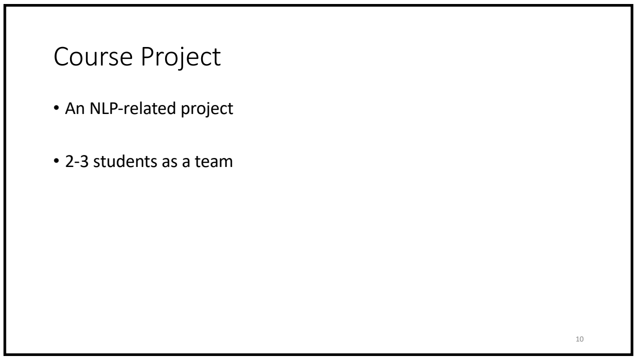
8



Grading

- Assignment (60%)
  - 4 assignments, 15% for each
- Project (35%) (details come up soon)
- Participation (5%)
  - Classes: attendance, ask and answer questions, participate in discussions…
  - Piazza: help your peers, address questions…

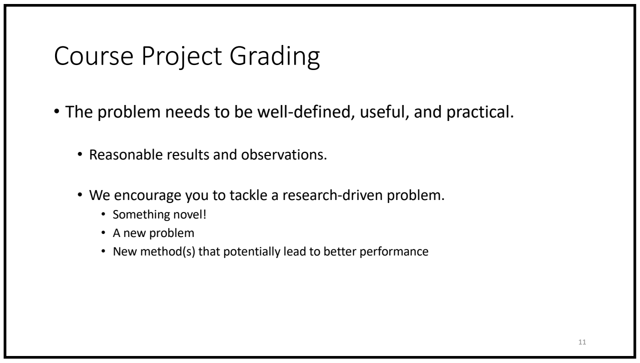
9



Course Project

- An NLP-related project
- 2-3 students as a team

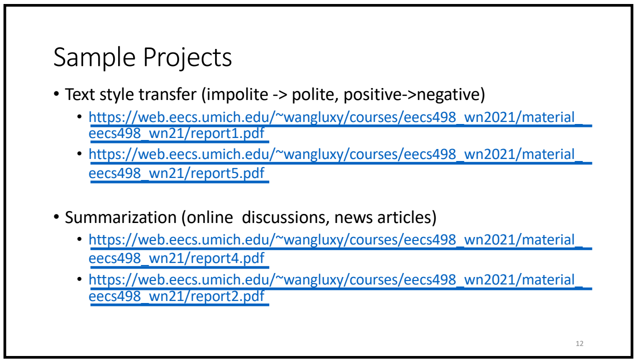
10



Course Project Grading

- The problem needs to be well-defined, useful, and practical.
  - Reasonable results and observations.
  - We encourage you to tackle a research-driven problem.
    - Something novel!
    - A new problem
    - New method(s) that potentially lead to better performance

11



Sample Projects

- Text style transfer (impolite -> polite, positive->negative)
  - https://web.eecs.umich.edu/~wangluxy/courses/eecs498_wn2021/material_eecs498_wn21/report1.pdf
  - https://web.eecs.umich.edu/~wangluxy/courses/eecs498_wn2021/material_eecs498_wn21/report5.pdf
- Summarization (online discussions, news articles)
  - https://web.eecs.umich.edu/~wangluxy/courses/eecs498_wn2021/material_eecs498_wn21/report4.pdf
  - https://web.eecs.umich.edu/~wangluxy/courses/eecs498_wn2021/material_eecs498_wn21/report2.pdf

12

More Project Samples

- Stanford NLP class
  - http://web.stanford.edu/class/cs224n
  - Notice its focus on deep learning
  - Your project can use any machine learning technique(s) on a natural language processing problem, and shouldn't be limited to deep learning only.

13

Course Project

- Talk to the instructor and IAs on project topics!
  - Zoom meetings (~10 minutes) will be arranged during the week of Feb 1st.
- How to find teammates?
  - Talk to your classmates and see if you share interests!
  - How to do it online: Post on piazza with your background (programming language and skills) + potential project ideas + your email contact, other students should feel free to reach out

14

Course Project Grading

- Three reports
  - One-page proposal (5%), due on Feb 12th at 11:59pm.
  - Progress report, with code (8%)
  - Final, with code (12%)
- One presentation
  - In class (7%)
  - feedback to other teams' presentations (3%)

15

Audience Award

- Bonus points!
  - All teams vote for their favorite project(s) after presentation.
  - The team gets the most votes will be awarded with 1% bonus point!

16

Submission and Late Policy

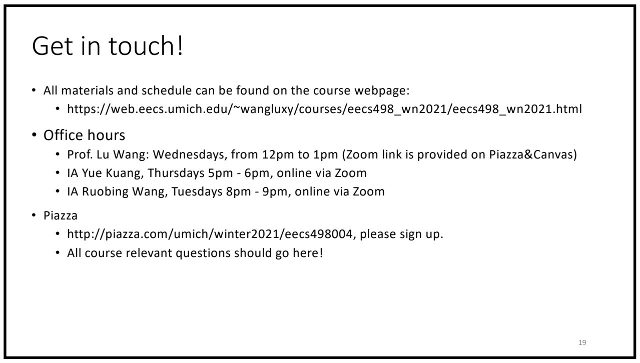
- Programming language
  - Python (recommended)
- All submissions are in electronic format.
  - Due on Canvas.

17

Submission and Late Policy

- Submissions turned in late will be charged 20 points (out of 100 points) off for each late day (i.e. 24 hours).
- Each student has a budget of **8 days in total** throughout the semester before a late penalty is applied.
- Late days are not applicable to presentations.
- Each group member is charged with the same number of late days, if any, for their submission.

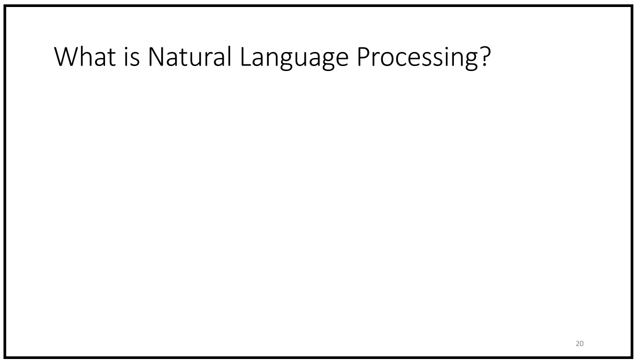
18



Get in touch!

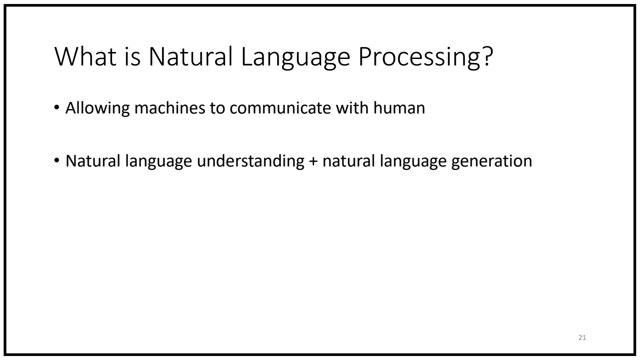
- All materials and schedule can be found on the course webpage:
  - https://web.eecs.umich.edu/~wangluxy/courses/eecs498_wn2021/eecs498_wn2021.html
- Office hours
  - Prof. Lu Wang: Wednesdays, from 12pm to 1pm (Zoom link is provided on Piazza&Canvas)
  - IA Yue Kuang, Thursdays 5pm - 6pm, online via Zoom
  - IA Ruobing Wang, Tuesdays 8pm - 9pm, online via Zoom
- Piazza
  - http://piazza.com/umich/winter2021/eecs498004, please sign up.
  - All course relevant questions should go here!

19



What is Natural Language Processing?

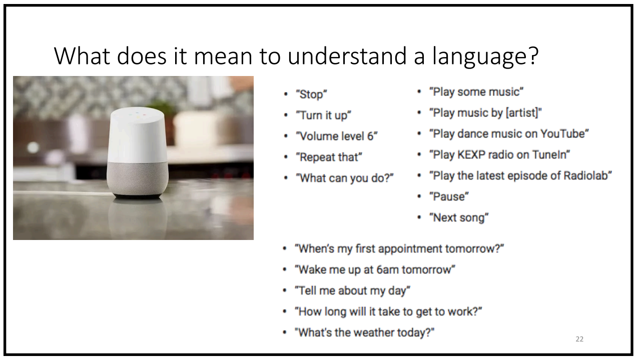
20



What is Natural Language Processing?

- Allowing machines to communicate with human
- Natural language understanding + natural language generation

21



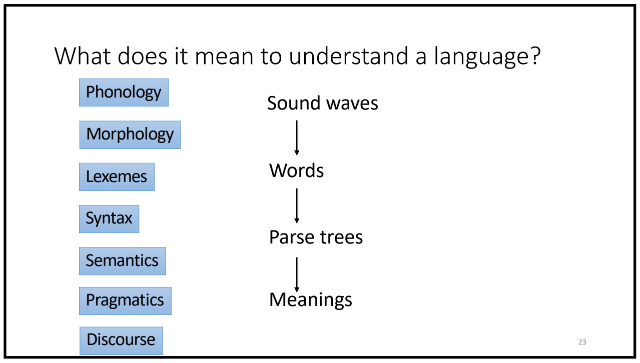
What does it mean to understand a language?

- "Stop"
- "Turn it up"
- "Volume level 6"
- "Repeat that"
- "What can you do?"

- "Play some music"
- "Play music by [artist]"
- "Play dance music on YouTube"
- "Play KEXP radio on TuneIn"
- "Play the latest episode of Radiolab"
- "Pause"
- "Next song"

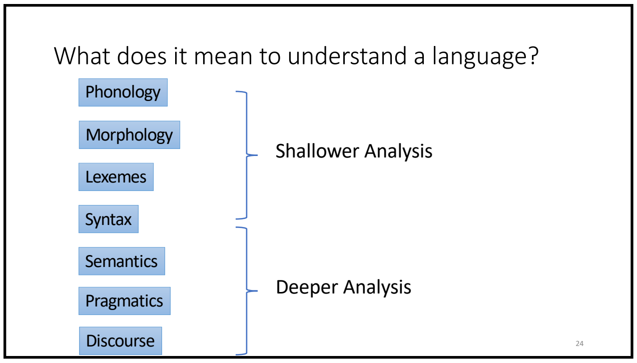
- "When's my first appointment tomorrow?"
- "Wake me up at 6am tomorrow"
- "Tell me about my day"
- "How long will it take to get to work?"
- "What's the weather today?"

22



What does it mean to understand a language?

23



What does it mean to understand a language?

24

Syntax, Semantics, Pragmatics

- Syntax concerns the proper ordering of words and its affect on meaning.
  - The dog bit the boy.
  - The boy bit the dog.
  - Bit boy dog the the.
- Semantics concerns the (literal) meaning of words, phrases, and sentences.
  - “plant” as a photosynthetic organism
  - “plant” as a manufacturing facility
  - “plant” as the act of sowing
- Pragmatics concerns the overall communicative and social context and its effect on interpretation.
  - Honest or dishonest?
  - Context 1: Kyle and Ellen would like to see a movie. Kyle has $20 in his pocket. Tickets cost $8 each.
  - Context 2: Kyle and Ellen would like to see a movie. Kyle has $20 in his pocket. Tickets cost $10 each.

25

Syntax, Semantics, Pragmatics

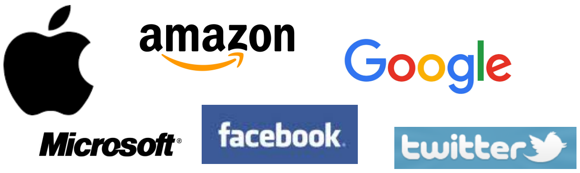
- Syntax concerns the proper ordering of words and its affect on meaning.
  - The dog bit the boy.
  - The boy bit the dog.
  - Bit boy dog the the.
- Semantics concerns the (literal) meaning of words, phrases, and sentences.
  - “plant” as a photosynthetic organism
  - “plant” as a manufacturing facility
  - “plant” as the act of sowing
- Pragmatics concerns the overall communicative and social context and its effect on interpretation.
  - Honest or dishonest?
  - Context 1: Kyle and Ellen would like to see a movie. Kyle has $20 in his pocket. Tickets cost $8 each.
  - Context 2: Kyle and Ellen would like to see a movie. Kyle has $20 in his pocket. Tickets cost $10 each.
  - Kyle: “I have $8.”

26

Where NLP is used?

27

Commercial World



28

Social World

- Disaster Relief
- Chatbots for Mental Health
- Detecting abusive language in online posts

29

Text Classification: Disaster Response

- Haiti Earthquake 2010
- Classifying SMS messages

Mwen thomassin 32 nan pyron mwen ta renmen jwen yon ti dlo gras a dieu bo lakay mwen anfom se sel dlo nou bezwen

> I am in Thomassin number 32, in the area named Pyron. I would like to have some water. Thank God we are fine, but we desperately need water.

30

Extracting Social Meaning from Language

- Uncertainty (students in tutoring)
- Annoyance (callers to dialogue systems)
- Anger (police-community interaction)
- Deception
- Emotion
- Intoxication
- Flirtation, Romantic interest

31

31

Sentiment in Restaurant Reviews

A very bad (one-star) review:

The bartender... absolutely horrible... we waited 10 min before we even got her attention... and then we had to wait 45 - FORTY FIVE! - minutes for our entrees... stalk the waitress to get the cheque... she didn't make eye contact or even break her stride to wait for a response ...

Dan Jurafsky, Victor Chahuneau, Bryan R. Routledge, and Noah A. Smith. 2014. Narrative framing of consumer sentiment in online restaurant reviews. First Monday 19:4

32

32

What is the language of bad reviews?

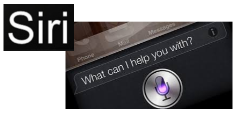
- Negative sentiment language
  horrible awful terrible bad disgusting
- Past narratives about people
  waited, didn't, was
  he, she, his, her,
  manager, customer, waitress, waiter
- Frequent mentions of we and us
  ... **we** were ignored until **we** flagged down a waiter to get **our** waitress ...

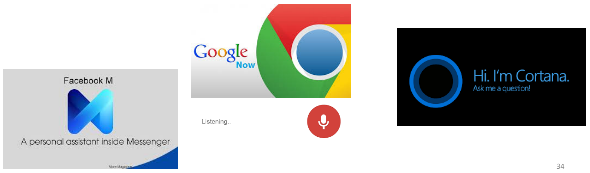
33

33

Personal Assistants

amazon alexa

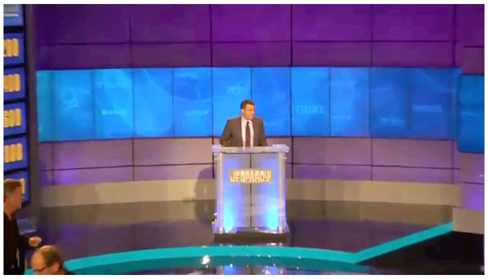




34

34

Question Answering: IBM's Watson

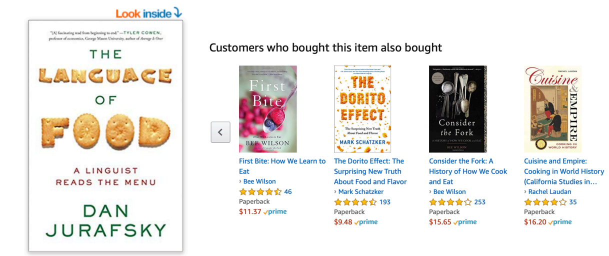


35

35

Recommendation Engines

If you bought....



Customers who bought this item also bought

First Bite: How We Learn to Eat
› Bee Wilson
★★★★☆ 46
Paperback
$11.37 ✓prime

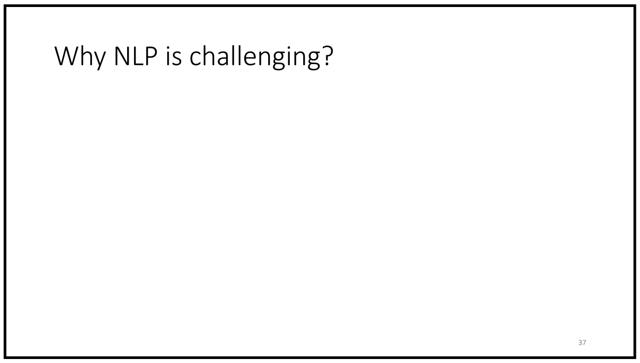
The Dorito Effect: The Surprising New Truth About Food and Flavor
› Mark Schatzker
★★★★☆ 193
Paperback
$9.48 ✓prime

Consider the Fork: A History of How We Cook and Eat
› Bee Wilson
★★★★☆ 253
Paperback
$15.65 ✓prime

Cuisine and Empire: Cooking in World History (California Studies in…
› Rachel Laudan
★★★★☆ 35
Paperback
$16.20 ✓prime

36

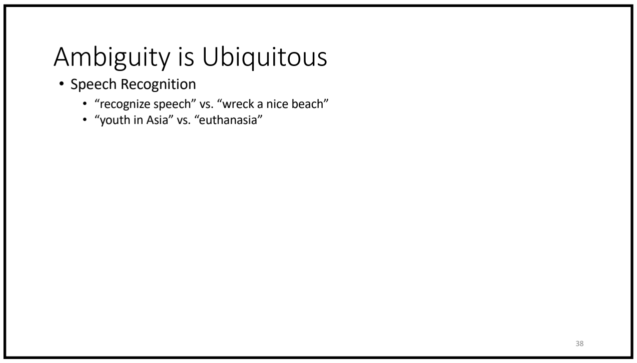
36



Why NLP is challenging?

37

37

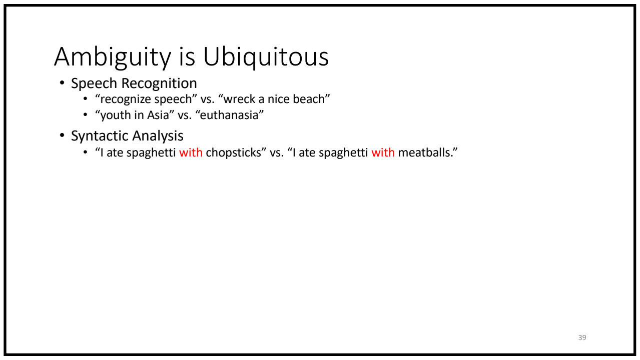


Ambiguity is Ubiquitous

- Speech Recognition
  - "recognize speech" vs. "wreck a nice beach"
  - "youth in Asia" vs. "euthanasia"

38

38

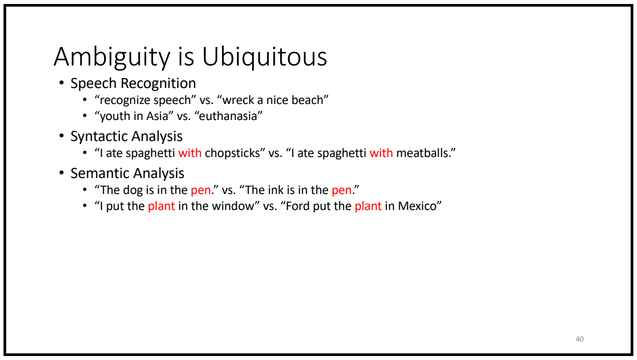


Ambiguity is Ubiquitous

- Speech Recognition
  - "recognize speech" vs. "wreck a nice beach"
  - "youth in Asia" vs. "euthanasia"
- Syntactic Analysis
  - "I ate spaghetti with chopsticks" vs. "I ate spaghetti with meatballs."

39

39

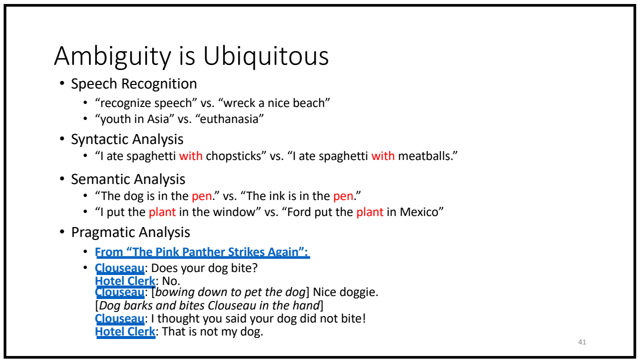


Ambiguity is Ubiquitous

- Speech Recognition
  - "recognize speech" vs. "wreck a nice beach"
  - "youth in Asia" vs. "euthanasia"
- Syntactic Analysis
  - "I ate spaghetti with chopsticks" vs. "I ate spaghetti with meatballs."
- Semantic Analysis
  - "The dog is in the pen." vs. "The ink is in the pen."
  - "I put the plant in the window" vs. "Ford put the plant in Mexico"

40

40

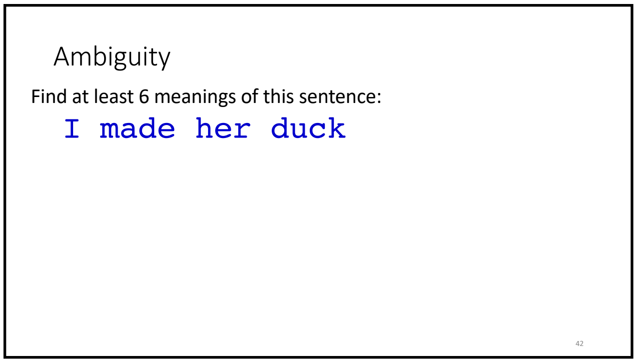


Ambiguity is Ubiquitous

- Speech Recognition
  - "recognize speech" vs. "wreck a nice beach"
  - "youth in Asia" vs. "euthanasia"
- Syntactic Analysis
  - "I ate spaghetti with chopsticks" vs. "I ate spaghetti with meatballs."
- Semantic Analysis
  - "The dog is in the pen." vs. "The ink is in the pen."
  - "I put the plant in the window" vs. "Ford put the plant in Mexico"
- Pragmatic Analysis
  - **From "The Pink Panther Strikes Again":**
  - **Clouseau**: Does your dog bite?
    Hotel Clerk: No.
    Clouseau: [*bowing down to pet the dog*] Nice doggie.
    [*Dog barks and bites Clouseau in the hand*]
    Clouseau: I thought you said your dog did not bite!
    Hotel Clerk: That is not my dog.

41

41



Ambiguity

Find at least 6 meanings of this sentence:

`I made her duck`

42

42

Ambiguity

Find at least 6 meanings of this sentence:

I made her duck

- I cooked waterfowl for her benefit (to eat)
- I cooked waterfowl belonging to her
- I created the (plaster?) waterfowl she owns
- I caused her to quickly lower her head or body
- I recognized the true identity of her spy waterfowl
- I waved my magic wand and turned her into undifferentiated waterfowl

43

Ambiguity

I caused her to quickly lower her head or body

Part of speech: “duck” can be a Noun or Verb

I cooked waterfowl belonging to her.

Part of speech:

“her” is possessive pronoun (“of her”)

“her” is dative pronoun (“for her”)

I made the (plaster) duck statue she owns

Word Meaning : “make” can mean “create” or “cook”

44

Ambiguity is Explosive

- Ambiguities compound to generate enormous numbers of possible interpretations.
- In English, a sentence ending in *n* prepositional phrases has *over* 2^n syntactic interpretations
  - ˝I saw the man with the telescope”: 2 parses
  - “I saw the man on the hill with the telescope.”: 5 parses
  - “I saw the man on the hill in Texas with the telescope”: 14 parses
  - “I saw the man on the hill in Texas with the telescope at noon.”: 42 parses
  - “I saw the man on the hill in Texas with the telescope at noon on Monday”: 132 parses

45

Humor and Ambiguity

- Many jokes rely on the ambiguity of language:
  - Policeman to little boy: “We are looking for a thief with a bicycle.” Little boy: “Wouldn’t you be better using your eyes.”
  - Why is the teacher wearing sun-glasses. Because the class is so bright.
  - Groucho Marx: One morning I shot an elephant in my pajamas. How he got into my pajamas, I’ll never know.
  - She criticized my apartment, so I knocked her flat.
  - Noah took all of the animals on the ark in pairs. Except the worms, they came in apples.

46

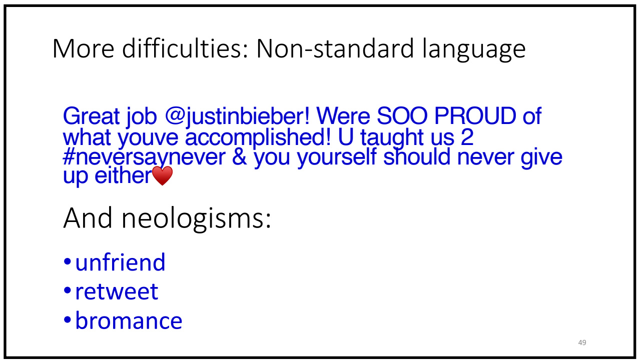
Why is Language Ambiguous?

47

Why is Language Ambiguous?

- Having a unique linguistic expression for every possible conceptualization that could be conveyed would make language overly complex and linguistic expressions unnecessarily long.
- Allowing resolvable ambiguity permits shorter linguistic expressions, i.e. data compression.
- Language relies on people’s ability to use their knowledge and inference abilities to properly resolve ambiguities.
- Infrequently, disambiguation fails, i.e. the compression is lossy.

48



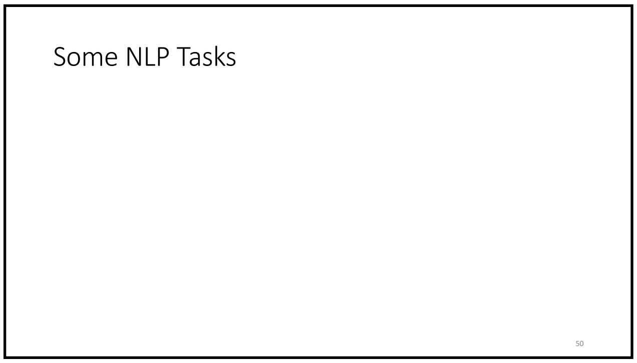
More difficulties: Non-standard language

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

And neologisms:

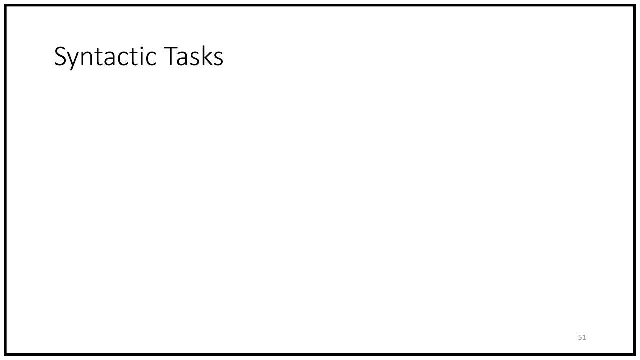
- unfriend
- retweet
- bromance

49



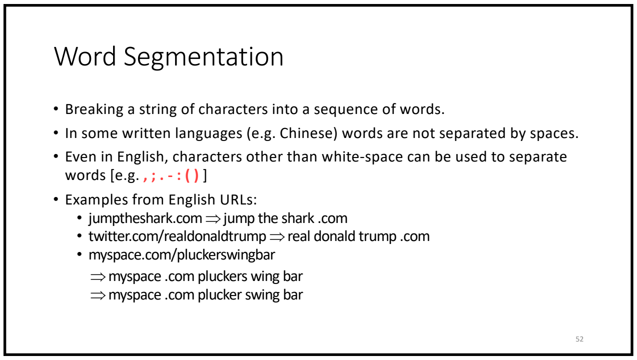
Some NLP Tasks

50



Syntactic Tasks

51



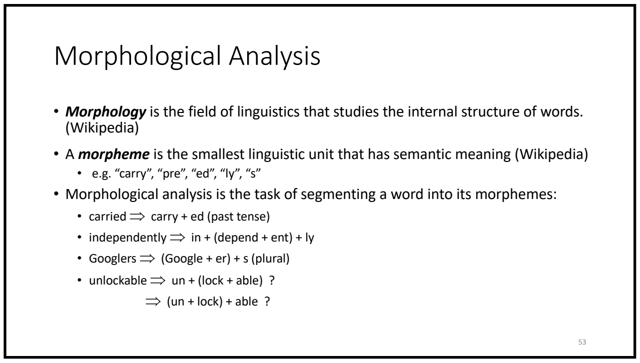
Word Segmentation

- Breaking a string of characters into a sequence of words.
- In some written languages (e.g. Chinese) words are not separated by spaces.
- Even in English, characters other than white-space can be used to separate words [e.g. , ; . - : ( )]
- Examples from English URLs:
  - jumptheshark.com ⇒ jump the shark .com
  - twitter.com/realdonaldtrump ⇒ real donald trump .com
  - myspace.com/pluckerswingbar

    ⇒ myspace .com pluckers wing bar

    ⇒ myspace .com plucker swing bar

52

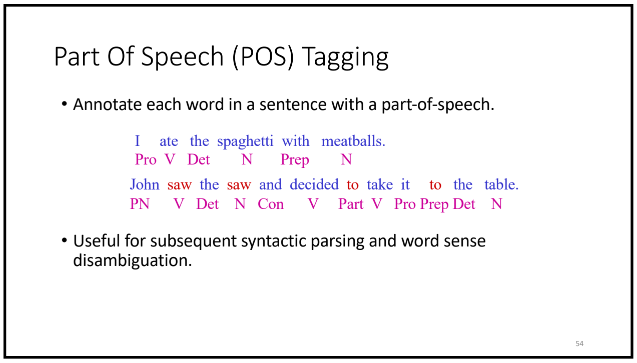


Morphological Analysis

- ***Morphology*** is the field of linguistics that studies the internal structure of words. (Wikipedia)
- A ***morpheme*** is the smallest linguistic unit that has semantic meaning (Wikipedia)
  - e.g. “carry”, “pre”, “ed”, “ly”, “s”
- Morphological analysis is the task of segmenting a word into its morphemes:
  - carried ⇒ carry + ed (past tense)
  - independently ⇒ in + (depend + ent) + ly
  - Googlers ⇒ (Google + er) + s (plural)
  - unlockable ⇒ un + (lock + able) ?

    ⇒ (un + lock) + able ?

53



Part Of Speech (POS) Tagging

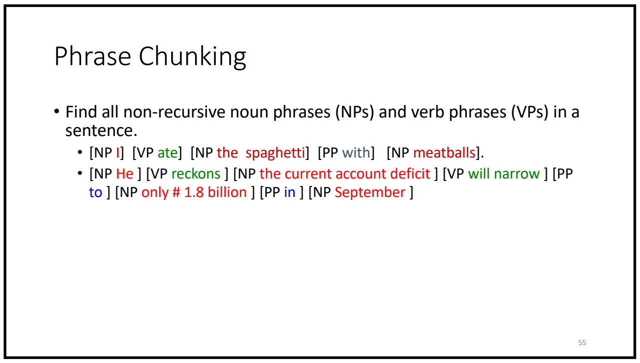
- Annotate each word in a sentence with a part-of-speech.

I ate the spaghetti with meatballs.
Pro V Det N Prep N

John saw the saw and decided to take it to the table.
PN V Det N Con V Part V Pro Prep Det N

- Useful for subsequent syntactic parsing and word sense disambiguation.

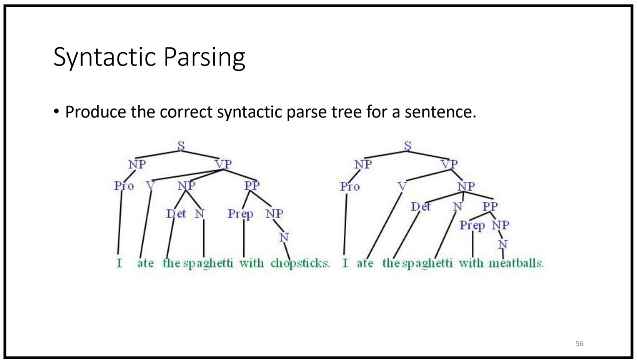
54



Phrase Chunking

- Find all non-recursive noun phrases (NPs) and verb phrases (VPs) in a sentence.
  - [NP I] [VP ate] [NP the spaghetti] [PP with] [NP meatballs].
  - [NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September]

55

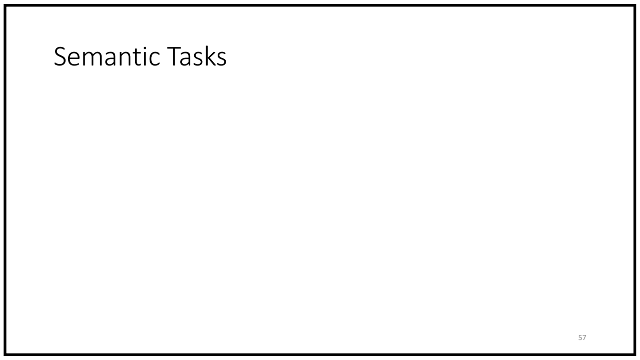


Syntactic Parsing

- Produce the correct syntactic parse tree for a sentence.

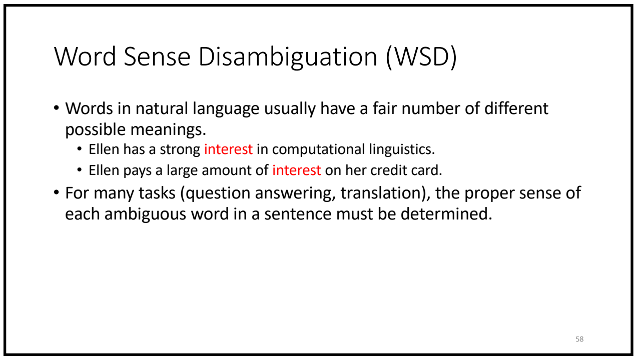
S
NP
VP
Pro
V
NP
PP
Det
N
Prep
NP
N
I ate the spaghetti with chopsticks.
S
NP
VP
Pro
V
NP
Det
N
PP
Prep
NP
N
I ate the spaghetti with meatballs.

56



Semantic Tasks

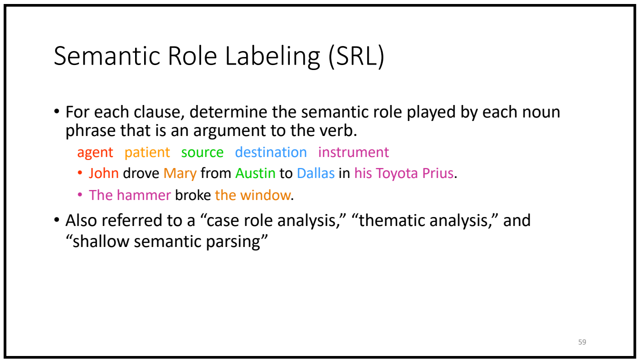
57



Word Sense Disambiguation (WSD)

- Words in natural language usually have a fair number of different possible meanings.
  - Ellen has a strong interest in computational linguistics.
  - Ellen pays a large amount of interest on her credit card.
- For many tasks (question answering, translation), the proper sense of each ambiguous word in a sentence must be determined.

58

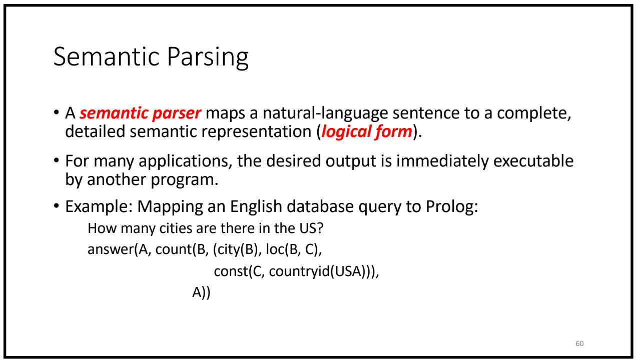


Semantic Role Labeling (SRL)

- For each clause, determine the semantic role played by each noun phrase that is an argument to the verb.

  agent patient source destination instrument
  - John drove Mary from Austin to Dallas in his Toyota Prius.
  - The hammer broke the window.
- Also referred to a “case role analysis,” “thematic analysis,” and “shallow semantic parsing”

59



Semantic Parsing

- A ***semantic parser*** maps a natural-language sentence to a complete, detailed semantic representation (***logical form***).
- For many applications, the desired output is immediately executable by another program.
- Example: Mapping an English database query to Prolog:

  How many cities are there in the US?

  answer(A, count(B, (city(B), loc(B, C),
  const(C, countryid(USA))),
  A))

60

Textual Entailment

- Determine whether one natural language sentence entails (implies) another under an ordinary interpretation.
- E.g., “A soccer game with multiple males playing. -> Some men are playing a sport.”

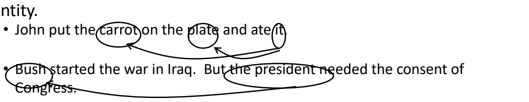
61

Pragmatics/Discourse Tasks

62

Anaphora Resolution/Co-Reference

- Determine which phrases in a document refer to the same underlying entity.



  - John put the carrot on the plate and ate it.
  - Bush started the war in Iraq. But the president needed the consent of Congress.
- Some cases require difficult reasoning.
  - Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. "Jack has a kite. He will make you take it back."

63

More Application-driven Tasks

64

Information Extraction (IE)

- Identify phrases in language that refer to specific types of entities and relations in text.
- Named entity recognition is task of identifying names of people, places, organizations, etc. in text.

  people organizations places

  - Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.
- Relation extraction identifies specific relations between entities.
  - Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.
  - Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.

65

Question Answering

- Directly answer natural language questions based on information presented in a corpora of textual documents (e.g. the web).
  - Who is the president of United States?
    - Donald Trump
  - What is the popular of Massachusetts?
    - 6.8 million

66

Text Summarization

- Produce a short summary of one or many longer document(s).
  - **Article:** An international team of scientists studied diet and mortality in 135,335 people between 35 and 70 years old in 18 countries, following them for an average of more than seven years. Diet information depended on self-reports, and the scientists controlled for factors including age, sex, smoking, physical activity and body mass index. The study is in The Lancet. Compared with people who ate the lowest 20 percent of carbohydrates, those who ate the highest 20 percent had a 28 percent increased risk of death. But high carbohydrate intake was not associated with cardiovascular death. …
  - **Summary:** Researchers found that people who ate higher amounts of carbohydrates had a higher risk of dying than those who ate more fats.

67

Spoken Dialogue Systems -- Chatbots

- Q: Is it going to rain today?
- A: It will be mostly sunny. No rain is expected.

68

Machine Translation

- Translate a sentence from one natural language to another.
  - 我喜欢汉堡 → I like burgers.

69

Ambiguity Resolution is Required for Translation

- Syntactic and semantic ambiguities must be properly resolved for correct translation:
  - "John plays the guitar." → "John 弹 吉他"
  - "John plays soccer." → "John 踢 足球"

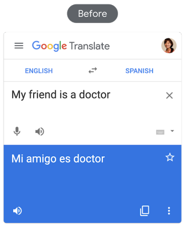
70

Ambiguity Resolution is Required for Translation

- Syntactic and semantic ambiguities must be properly resolved for correct translation:
  - "John plays the guitar." → "John 弹 吉他"
  - "John plays soccer." → "John 踢 足球"
- An apocryphal story is that an early MT system gave the following results when translating from English to Russian and then back to English:
  - "The spirit is willing but the flesh is weak." → "The liquor is good but the meat is spoiled."
  - "Out of sight, out of mind." → "Invisible idiot."

71

Bias and Ethics

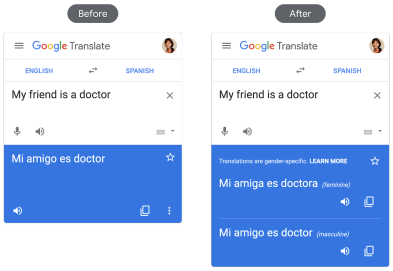


My friend is a doctor

Mi amigo es doctor

72

Bias and Ethics



Before

My friend is a doctor

Mi amigo es doctor

After

My friend is a doctor

Translations are gender-specific. LEARN MORE

Mi amiga es doctora *(feminine)*

Mi amigo es doctor *(masculine)*

73

Resolving Ambiguity

- Choosing the correct interpretation of linguistic utterances requires (commonsense) knowledge of:
  - Syntax
    - An agent is typically the subject of the verb
  - Semantics
    - Michael and Ellen are names of people
    - August is the name of a month (and of a person)
    - Toyota is a car company and Prius is a brand of car
  - Pragmatics
    - Some social norm, communicative goals
    - Asking a question, expecting an answer
  - World knowledge
    - Credit cards require users to pay financial interest
    - Agents must be animate and a hammer is not animate

74

State-of-the-Arts

- Learning from large amounts of text data (cf. rule-based methods)
  - Supervised learning or unsupervised learning
- Statistical machine learning-based methods
  - The probabilistic knowledge acquired allows robust processing that handles linguistic regularities as well as exceptions.
- Now with neural network-based methods mostly

75

Related Fields

- Artificial Intelligence
- Machine Learning
- Linguistics
- Cognitive science
- Logic
- Data science
- Political science
- Education
- Economics
- …many more

76

Relevant Scientific Conferences and Journals

- Association for Computational Linguistics (ACL)
- North American Association for Computational Linguistics (NAACL)
- Empirical Methods in Natural Language Processing (EMNLP)
- International Conference on Computational Linguistics (COLING)
- Conference on Computational Natural Language Learning (CoNLL)
- Transactions of the Association for Computational Linguistics (TACL)
- Journal of Computational Linguistics (CL)

77