

## CMPT 413/713: Natural Language Processing

Angel Xuan Chang 2024-01-08

Adapted from slides from Anoop Sarkar, Danqi Chen and Karthik Narasimhan

1

## Today

### • Intro to NLP

- What is language? NLP vs computational linguistics.
- Brief history of NLP
- Levels of representation in language
- Key questions of NLP
- Course Logistics

## NLP is everywhere





Google translate

### Virtual assistants

## Information finding

Google

when were the first pyramids built

**୍ର** ପ

🔍 All 🖾 Images 🗉 News 🔗 Shopping 🕩 Videos : More Settings Tools

About 20,300,000 results (0.67 seconds)

### 2630 BC

Most were built as tombs for the country's pharaohs and their consorts during the Old and Middle Kingdom periods. The earliest known Egyptian pyramids are found at Saqqara, northwest of Memphis. The earliest among these is the Pyramid of Djoser (constructed **2630 BC**-2611 BC) which was built during the third dynasty.

The First Pyramids Built - Timeline Index www.timelineindex.com > content > view

# Text to image generation

a painting of a fox sitting in a field at sunrise in the style of Claude Monet



https://openai.com/dall-e-2/



A cute sloth holding a small treasure chest. A bright golden glow is coming from the chest.



https://imagen.research.google/

Try it out yourself: <u>https://huggingface.co/spaces/stabilityai/stable-diffusion</u>, <u>https://www.craiyon.com/</u>, <u>https://www.midjourney.com/home/</u>

## Copilot

### Generate code from text!

```
⊤s sentiments.ts
                              🇬 parse_expenses.py
                                                  🛃 addresses.rb
                🗠 write_sql.go
 1 #!/usr/bin/env ts-node
 3 import { fetch } from "fetch-h2";
 5 // Determine whether the sentiment of text is positive
6 // Use a web service
 7 async function isPositive(text: string): Promise<boolean> {
     const response = await fetch(`http://text-processing.com/api/sentiment/`, {
       method: "POST",
       body: `text=${text}`,
11
       headers: {
12
         "Content-Type": "application/x-www-form-urlencoded",
       },
     });
     const json = await response.json();
     return json.label === "pos";
   🔠 Copilot
```

https://github.com/features/copilot

- Language is used to communicate
  - Things, actions, abstract concepts





- Language puts categories on the world
  - It discretizes a continuous space





- Language picks out what is salient and important
- What concepts do we have words for?



- Language picks out what is salient and important
- What concepts do we have words for?
- Different languages have different discretization boundaries



- Language picks out what is salient and important
- What concepts do we have words for?
- Different languages have different discretization boundaries



http://pyersqr.org/classes/Ling731/Space2.htm

## Natural Language Processing



## Building useful system to process language

## **Computational Linguistics**



(image credit: https://www.enterrasolutions.com/blog/computational-linguistics-and-natural-language-processing/)

## Using computers to study human language

## Analyzing word usage in literature



"The Transformation of Gender in English-Language Fiction," Cultural Analytics

## Brief history of NLP

## Beginnings

This card is punc interpreted at the then accepted by translated by mea programs into the	hed with a top) in st the 701, c ans of stor	andard IBM punc onverted into its ed dictionary an language equival	I III II language sentence (as hed-card code. It is own binary language a d operational syntactic ent which is thep printy	1 10 1111110 1 1 1 1 1 1 1 1 1 1 1 1 1 1	Georgetown- IBM experiment.
"	',	`	·		1954
THE QUALITY CALORY CONTENT	OF	COAL	IS DETERMINED	BY	

Specimen punched card and below a strip with translation, printed within a few seconds

"Within three or five years, machine translation will be a solved problem"

## SHRDLU (Winograd, 1968)

Video of actual system: <a href="https://www.youtube.com/watch?v=bo4RvYJYOzI">https://www.youtube.com/watch?v=bo4RvYJYOzI</a>

Person: Pick up a big red block.

Computer: OK.

Person: Grasp the pyramid.

Computer: I don't understand which pyramid you mean.

Person (changing their mind): Find a block which is taller than the one you are holding and put it into the box.

Computer: By "it", I assume you mean the block which is taller than the one I am holding.

Computer: OK.

Person: What does the box contain?

Computer: The blue pyramid and the blue block.

Person: What is the pyramid supported by?

Computer: The box.







\*.ed => past tense verb Contained keyword: delicious

## Eliza (Weizenbaum, 1966)

Demo: <a href="http://psych.fullerton.edu/mbirnbaum/psych101/Eliza.htm">http://psych.fullerton.edu/mbirnbaum/psych101/Eliza.htm</a>

Men are all alike. IN WHAT WAY

They're always bugging us about something or other. CAN YOU THINK OF A SPECIFIC EXAMPLE

Well, my boyfriend made me come here. YOUR BOYFRIEND MADE YOU COME HERE

He says I'm depressed much of the time. I AM SORRY TO HEAR YOU ARE DEPRESSED Rogerian psychologist: reflect back what the patient said

- Set of ranked transformation rules based on keywords
- Apply rules based on matching words in sentence

## Transformation rules

Keywords are ranked from specific to general

- I know everybody laughed at me
- "I" is a very general keyword:
   I: (I \*) -> (You say you 2)
   YOU SAY YOU KNOW EVERYBODY LAUGHED AT YOU
- "Everybody" is much more interesting (someone using universals like everybody/always is probably "referring to some quite specific event or person")

WHO IN PARTICULAR ARE YOU THINKING OF?

- Implementation: keywords stored with their rank
   Everybody 5 (transformation rules)
  - **I 0** (transformation rules)

Backoff Please go on That's very interesting I see

(Adapted from slides: Stanford CS124N, Dan Jurafsky)



Rosie from the Jetsons



# language is hard!

Understanding

The Far Side - Gary Larson

## Why is natural language processing hard?

### Programming Languages C, C++, Java, Python, ...

- Unambiguous
- Fixed
- Designed
- Learnable?
- Known simple semantics

Natural Languages French, English, Korean, Chinese, Tagalog, ...

- Ambiguous
- Evolving
- Transmitted
- Learnable
- Complex semantics

## Why is NLP hard?

Interpretation of language assumes a common basis of **world knowledge** and **context** 

- Ambiguous:
  - "bank", "bat"
  - "Milk Drinkers Turn to Powder"
- Variation: Many ways to say same the
- Context dependent:
  - natural language is under-specified





## Context-dependence

"I put the bowl on the **table**"



### "The numbers in the **table** don't add up"

Class (I)	Expected numbers		0 - E		(0-E) <sup>2</sup> /E	
	A (2)	B (3)	A (4)	B (5)	A (6)	B (7)
1	11.80	10.20	5.20	-5.20	2.292	2.651
11.	24.67	21.33	0.33	-0.33	0.004	0.005
11	39.15	33.85	-0.15	0.15	0.001	0.001
IV	48.81	42.19	-6.81	6.81	0.950	1.009
V	30.57	26.43	1.43	-1.43	0.067	0.077
Total	30.57	134.00	0	0	3314	3.833

 $g^2 = 3.314 + 3.033 = 7.147$ ,  $df = 4.0.10 \le P \le 0.50$ .

### "I made her duck"

I cooked her duck (the waterfowl) for dinner

I cooked the duck (the waterfowl) belonging to her.

I created a model of a duck that she now has.

I caused her to lower her head

## Some language humor

(from real newspaper headlines)

Kids make nutritious snacks

Stolen painting found by tree

Miners refuse to work after death

Squad helps dog bite victim

Killer sentenced to die for second time in 10 years

Lack of brains hinders research

Pooches and people digging newly opened park

## Using rules for NLP

- Translation with rules started in the 1950s
- Eliza chatbot (1966)
- SHRDLU grounded language understanding (1968)



## Rule-based Machine Translation

- Georgetown IBM translator
- Bilingual dictionary to map Russian words to their English counterparts
- But what if the word order is different? How to handle all the nuances of meaning?



1 minute video showing 1954 MT:

https://youtu.be/K-HfpsHPmvw

https://www.ibm.com/ibm/history/exhibits/701/701\_translator.html

Coming up rules is hard!

Let's learn from data!



https://christophm.github.io/interpretable-ml-book/terminology.html

## Rise of statistical learning

(late) 1980s to 2000s: Statistical MT, 2000s-2014: Statistical Phrase-Based MT



- Use of machine learning techniques in NLP
- Increase in computational capabilities
- Availability of electronic corpora
- Data-driven evaluation of models

## Rise of statistical learning

(late) 1980s to 2000s: Statistical MT, 2000s-2014: Statistical Phrase-Based MT



IBM Models for translation



Speech recognition

Anytime a linguist leaves the group the (speech) recognition rate goes up - Fred Jelinek

## Statistical Machine Translation

• Required a parallel corpora



## Rise of statistical learning

Required human-designed representations and input features

Var	Definition	Value in Fig. 5.2
$x_1$	$count(positive lexicon) \in doc)$	3
$x_2$	$count(negative \ lexicon) \in doc)$	2
<i>x</i> <sub>3</sub>	$\begin{cases} 1 & \text{if "no"} \in \text{doc} \\ 0 & \text{otherwise} \end{cases}$	1
$x_4$	$count(1st and 2nd pronouns \in doc)$	3
<i>x</i> <sub>5</sub>	$\begin{cases} 1 & \text{if "!"} \in \text{doc} \\ 0 & \text{otherwise} \end{cases}$	0
$x_6$	log(word count of doc)	$\ln(64) = 4.15$

### Can these features be automatically learned?

## Promise of deep learning

• Most NLP works in the past focused on humandesigned representations and input features

Var	Definition	Value in Fig. 5.2
$x_1$	$count(positive lexicon) \in doc)$	3
$x_2$	$count(negative lexicon) \in doc)$	2
<i>x</i> <sub>3</sub>	$\begin{cases} 1 & \text{if "no"} \in \text{doc} \\ 0 & \text{otherwise} \end{cases}$	1
$x_4$	$count(1st and 2nd pronouns \in doc)$	3
<i>x</i> <sub>5</sub>	$\begin{cases} 1 & \text{if "!"} \in \text{doc} \\ 0 & \text{otherwise} \end{cases}$	0
$x_6$	log(word count of doc)	$\ln(64) = 4.15$

- Representation learning attempts to automatically learn good features and representations
- Deep learning attempts to learn multiple levels of representation on increasing complexity/abstraction

## Deep learning era

2014-Present: Neural Machine Translation

• Led to significant advances in core NLP technologies





## Deep learning era

- Significant advances in core NLP technologies
- Essential ingredient: large-scale supervision, lots of compute
- Reduced manual effort less/zero feature engineering



## Machine Translation

BBC Sign in	选项 (英文)	检索	Q
NEWS   中文			繁
主页   国际   两岸   英国   评论	科技   财经   图辑   音频材料   视频材料   BBC英伦网		

#### 巴拿马首任驻华大使专访: 与台湾断交之后

他说,不认同中国"买走"合湾邦交国 的说法,与中国建立关系对巴拿马有 利,不担心影响与美国关系。

① 1小时前

巴拿马外交转向周年 中美大国博奕内 幕解密

尼加拉瓜运河成谜:人走楼空 "不再 提及"

触发萨尔瓦多与台湾断交的港口令美 国担忧



000#VISITPanamá

特别推荐



人民币贬值、楼价高涨: 消费降级中 产失去的优质生活

广告



观点:高铁"一地两 检"——中国的强势 港人 的无力

#### 江沂:两地价值观和制度差异巨 大,冲突无可避免。香港没有反对 的权利, 也没有反对的能力。



娱乐业的寒蝉效应

分,最低分范冰冰为0分。

() 2小时前

球 ① 2018年9月11日

> 日本人脚踹慰安妇铜 像引爆台湾人抗议

中蒙参加俄罗斯军演

"中俄靠拢论"再吸眼

日前发布的《中国影视明星社会责 <sup>① 2018年9月11日</sup> 任研究报告》中,最高分徐峥为78

大空望远错探测比"三



台湾民间发起东京奥运"正名"公投的 意义



## Machine Translation



#### Interview with Panama's first ambassador to China: After breaking diplomatic relations with Taiwan

He said that he does not agree with China's saying that "buy" Taiwan's diplomatic relations with China. Establishing relations with China is beneficial to Panama and does not worry about affecting relations with the United States.

I hour ago

Panamanian Diplomacy Turns to Anniversary

The Nicaragua Canal is a mystery: people go to the floor and "no longer mention"

The port that triggered the break of El Salvador and Taiwan has worried the United States



000 #VISITPanamá

#### Special recommendation



Renminbi depreciation, property prices are rising: consumption downgrades the loss of quality life in the middle class

ADVERTISING



The significance of the Taiwanese people's "referred to" referendum

China and Mongolia participate in the Russian military

内权利,也没有反对的能力。

#### Hussian m

**RESEARCH > PUBLICATIONS >** 

### Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html

Input sentence:	Translation (PBMT):	Translation (GNMT):	Translation (human):
李克強此行將啟動中加 總理年度對話機制,與 加拿大總理杜魯多舉行 兩國總理首次年度對 話。	Li Keqiang premier added this line to start the annual dialogue mechanism with the Canadian Prime Minister Trudeau two prime ministers held its first annual session.	Li Keqiang will start the annual dialogue mechanism with Prime Minister Trudeau of Canada and hold the first annual dialogue between the two premiers.	Li Keqiang will initiate the annual dialogue mechanism between premiers of China and Canada during this visit, and hold the first annual dialogue with Premier Trudeau of Canada.

(Wu et al., 2016)

**RESEARCH > PUBLICATIONS >** 

### Google's Neural Machine **Translation System: Bridging** the Gap between Human and **Machine Translation**

https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html

Table 10: Mean	of side-by-	side score	s on prod	uction data
	PBMT	GNMT	Human	Relative
				Improvement
$English \rightarrow Spanish$	4.885	5.428	5.504	87%
$\mathrm{English} \to \mathrm{French}$	4.932	5.295	5.496	64%
$\mathrm{English} \to \mathrm{Chinese}$	4.035	4.594	4.987	58%
$\mathrm{Spanish} \to \mathrm{English}$	4.872	5.187	5.372	63%
$\mathrm{French} \to \mathrm{English}$	5.046	5.343	5.404	83%
$\mathbf{Chinese} \to \mathbf{English}$	3.694	4.263	4.636	60%

(Wu et al., 2016)

## Machine Translation (across 100 languages)

120



https://www.microsoft.com/en-us/research/blog/microsoft-translator-now-translating-100-languages-and-counting

## Multimodal models



https://github.com/OFA-Sys/OFA

OFA (Wang et al, 2022)

## Power of Data

### CleverBot (2010)

User:	Do you like Doctor Who?
Cleverbot:	Yes, so funny.
User:	What's your favorite show?
Cleverbot:	I don't have one.
User:	What's your name?
Cleverbot:	Jason I already told you.
User:	No your real name.
Cleverbot:	Stephanie IS my real name.
User:	I thought it was Jason?
Cleverbot:	How did you come to that conclusion
User:	Because you said so.
Cleverbot:	I have to do my homework.

### https://www.cleverbot.com/

### How it works:

- Corpus of conversational turns
- Find the most similar sentence and copy the response
- Learn from human input

### What do you get?

- Something that someone say
- Incoherent conversation

## Power of Data

### Meena (Google, 2020)

**Conversation D** 

Human: I like how cute cows are

Meena: Ah, interesting. You like how they are cute, right?

Human: yeah, and they're realy smart

Meena: I heard they go to college

Human: cows go to college?

**Meena**: *I heard that a cow went to Harvard.* 

**Human**: what did the cow study?

Meena: Bovine sciences.

Human: do horses go to Harvard?

Meena: Horses go to Hayvard.

**Human**: that's a pretty good joke, I feel like you led me into it

Meena: You were trying to steer it elsewhere, I can see it.

How it works:

- Corpus of conversational turns (over 40B words)
- Train huge neural network
   (2.6 billion parameters) for
   30 days on 2048 TPUs cores
- Predict response given a sentence

### https://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html

## Power of Data LaMDA (Google, 2022)



### How it works:

- Pretrain on even more data
- Fine-tune so it has certain characteristics



### Is this chatbot sentient?

https://ai.googleblog.com/2022/01/lamda-towards-safe-grounded-and-high.html

https://youtu.be/CvD9U\_3hBPo

### ChatGPT From OpenAI Is A Bot Taking The Tech World By Storm



## Power of Data ChatGPT (OpenAI, 2022)

https://openai.com/blog/chatgpt/

write a sonnet about pokemon

A world of wonder, filled with creatures wild,
 A realm of magic, where adventures start,
 From Kanto to Kalos, far and wide,
 The world of Pokémon, alive and heart.

With Pikachu as guide, we set out to roam, To catch and train them all, to be the best, Battling Gym Leaders, overcoming foes, Our bond with creatures strong, a bond we'll never test.

But it's not just battles that define, The world of Pokémon, so full of cheer, It's friendships formed, and lessons learned, A world that's full of love, and never fear.

So let's go on a journey, you and I, Through the world of Pokémon, beneath the sky.

### How it works:

- Trained on conversational data
- Uses reinforcement learning with human feedback

## The Brilliance and Weirdness of ChatGPT

A new chatbot from OpenAI is inspiring awe, fear, stunts and attempts to circumvent its guardrails.

https://www.nytimes.com/2022/12/05/technology/ chatgpt-ai-twitter.html



#### Exam results (ordered by GPT-3.5 performance)

Estimated percentile lower bound (among test takers)







## NLP before deep learning



The big stages of NLP before the deep learning era.

https://medium.com/@antoine.louis/a-brief-history-of-natural-language-processing-part-1-ffbcb937ebce

## NLP in the deep learning era



The big stages of NLP in the deep learning era.

https://medium.com/@antoine.louis/a-brief-history-of-natural-language-processing-part-2-f5e575e8e37

## Many remaining challenges in NLP!

## Challenges in modern NLP

- Scale: Large number of phenomena
- Sparsity: Text data is often heavy-tailed



## Ambiguity in grounding

## I saw her duck



## Challenges in modern NLP

### • Bias: Models learn biases in available data



• Context: Knowledge bases, perception, interaction

Compositional generalization



Compositionality



Generalization

vap bo

## Challenges of understanding language

- Does solving language require solving AI?
- Language derives its meaning from use



'block' 'pillar' 'slab' 'beam'.



• To truly understand language: need **context**, **world knowledge**, **mental model** of other people