

# Introduction to Natural Language Processing

2/2565: FRA501 Introduction to Natural Language Processing with Deep learning  
Week 01

Paisit Khanarsa, Ph.D.

Institute of Field Robotics (FIBO), King Mongkut's University of Technology Thonburi

# Outlines

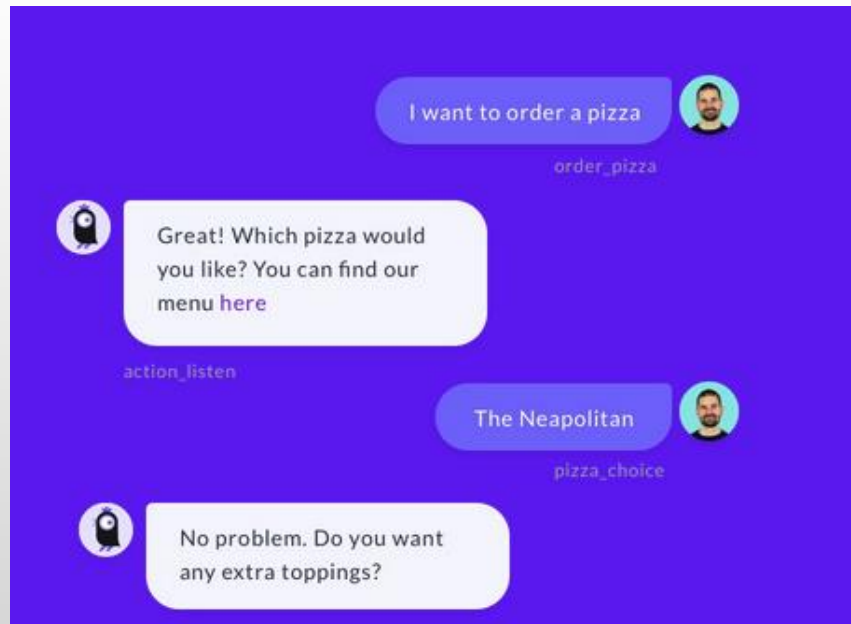
- What is Natural Language Processing (NLP) ?
- Why it is difficult ?
- Processing steps of NLP
- Deep Learning in NLP
- NLP tools

What is Natural Language Processing (NLP) ?

# Natural Language Processing (NLP)

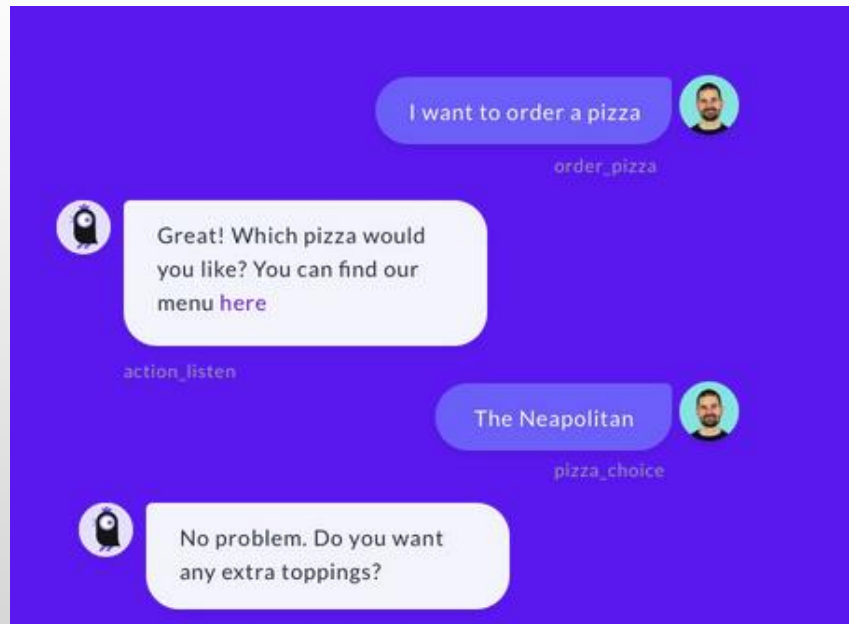
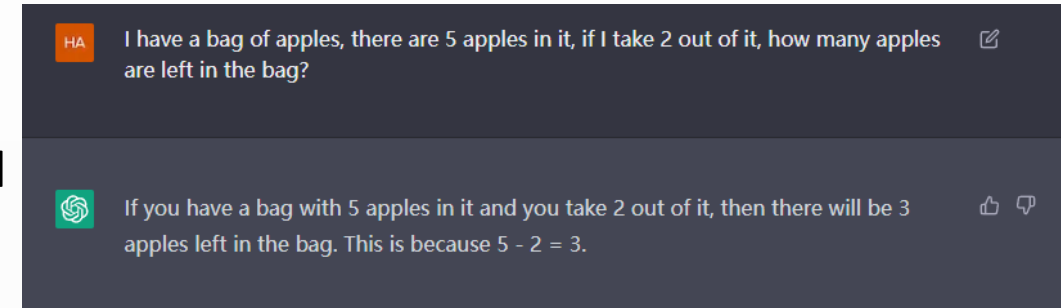
- NLP is a subfield in AI, where the goal is
  - To bridge the gap between how people communicate and what machines understand in order to perform useful tasks, e.g. making appointments, translation, question answering (QA), etc.

HA I have a bag of apples, there are 5 apples in it, if I take 2 out of it, how many apples are left in the bag? ✎



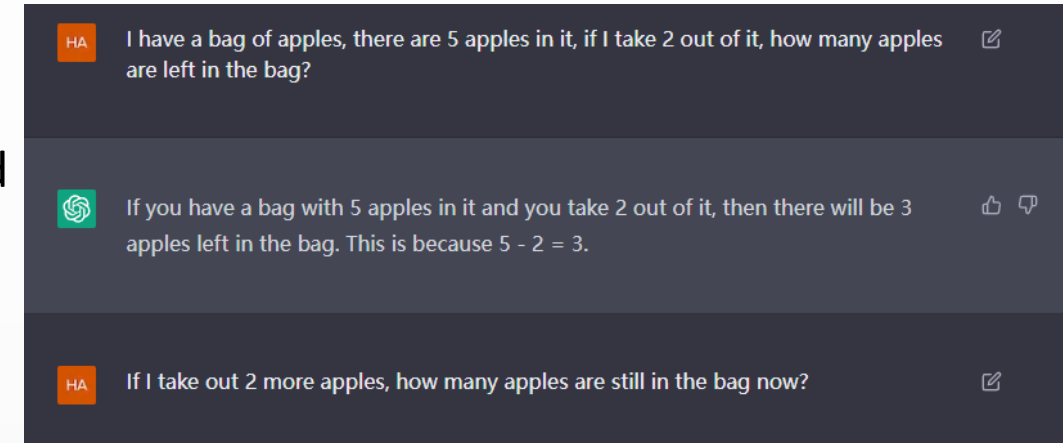
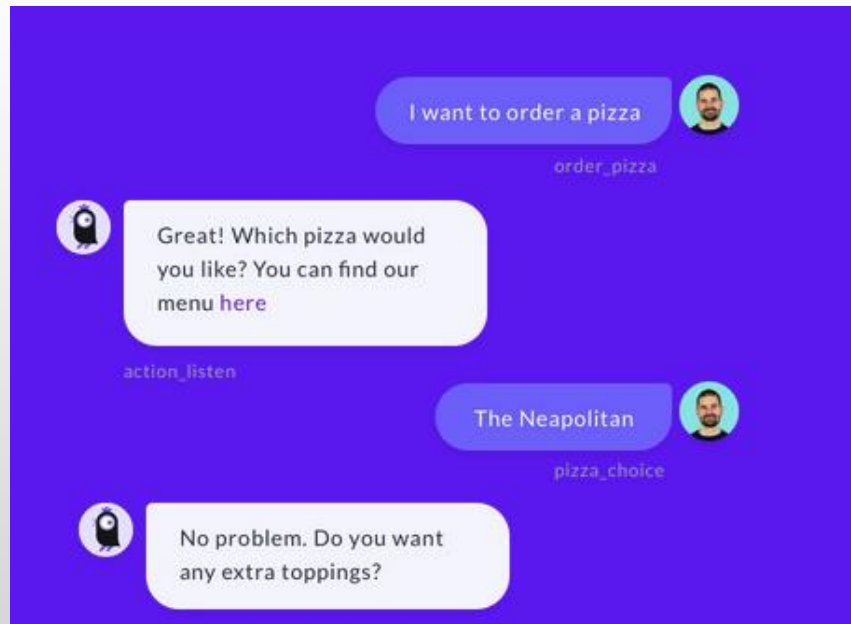
# Natural Language Processing (NLP)

- NLP is a subfield in AI, where the goal is
  - To bridge the gap between how people communicate and what machines understand in order to perform useful tasks, e.g. making appointments, translation, question answering (QA), etc.



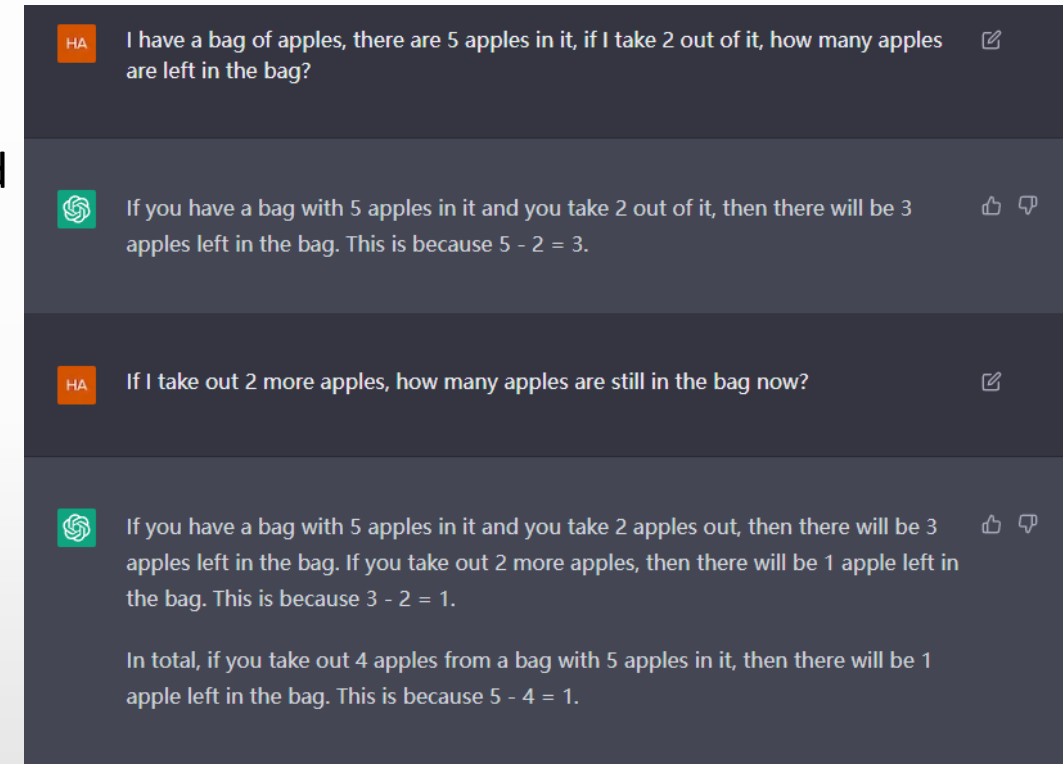
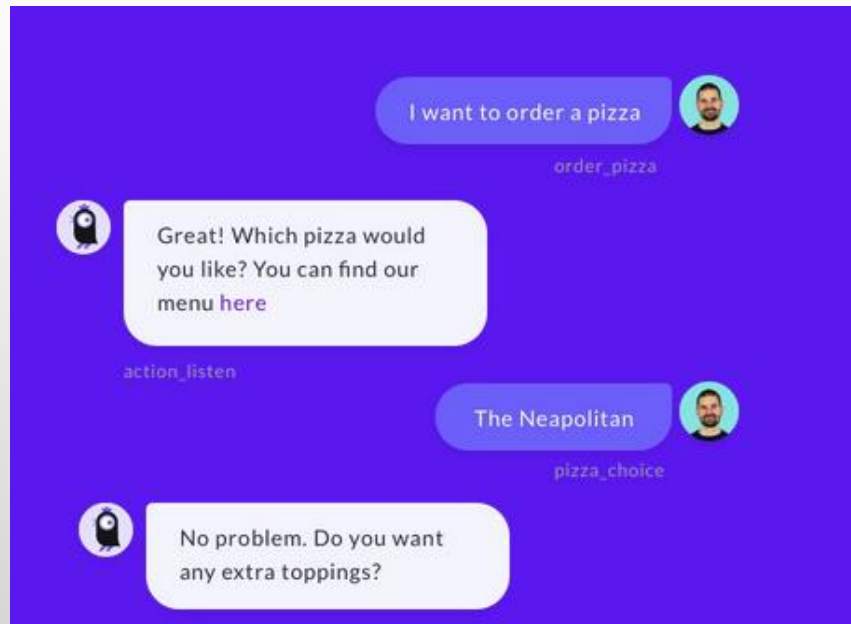
# Natural Language Processing (NLP)

- NLP is a subfield in AI, where the goal is
  - To bridge the gap between how people communicate and what machines understand in order to perform useful tasks, e.g. making appointments, translation, question answering (QA), etc.



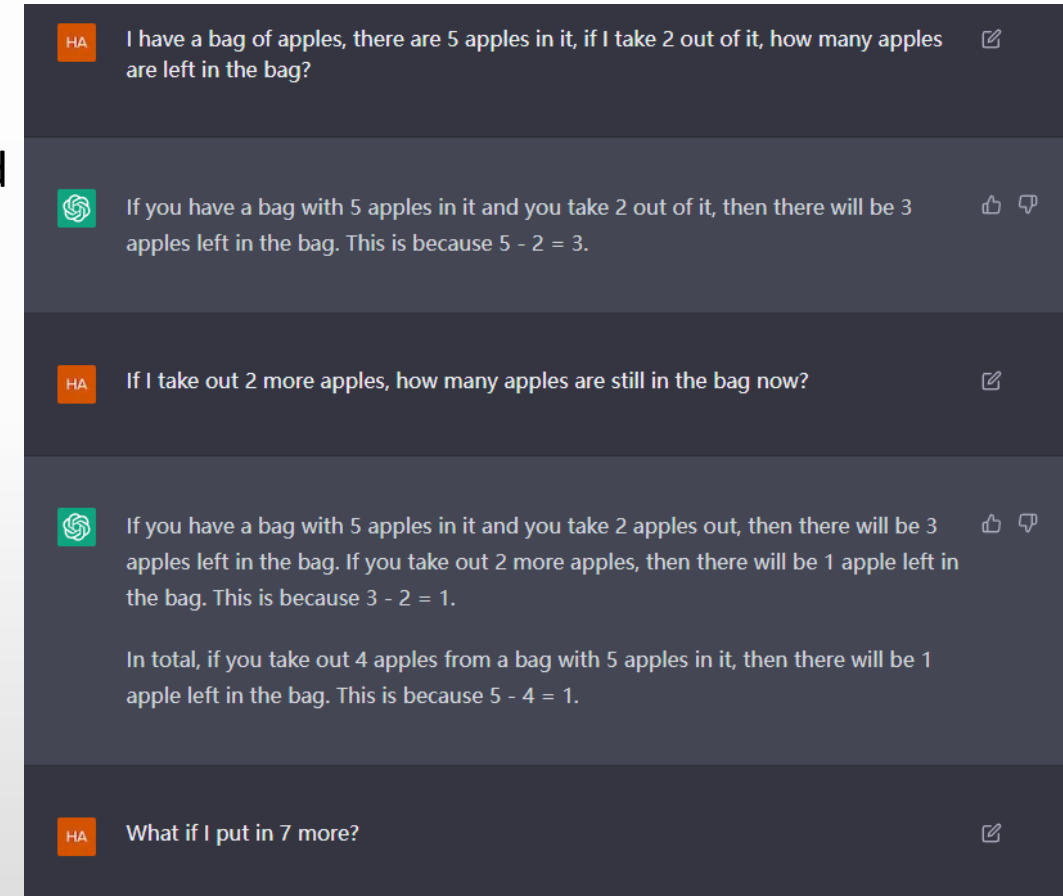
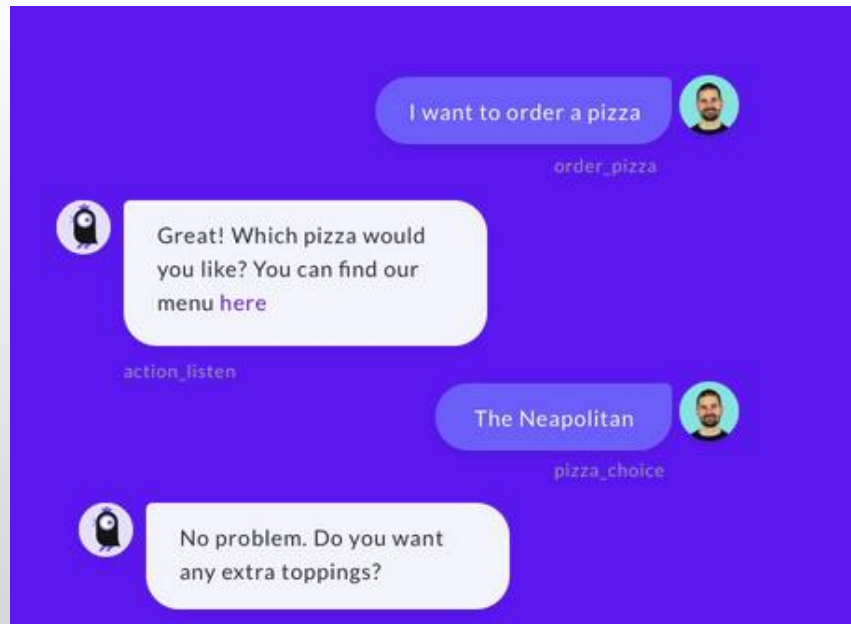
# Natural Language Processing (NLP)

- NLP is a subfield in AI, where the goal is
  - To bridge the gap between how people communicate and what machines understand in order to perform useful tasks, e.g. making appointments, translation, question answering (QA), etc.



# Natural Language Processing (NLP)

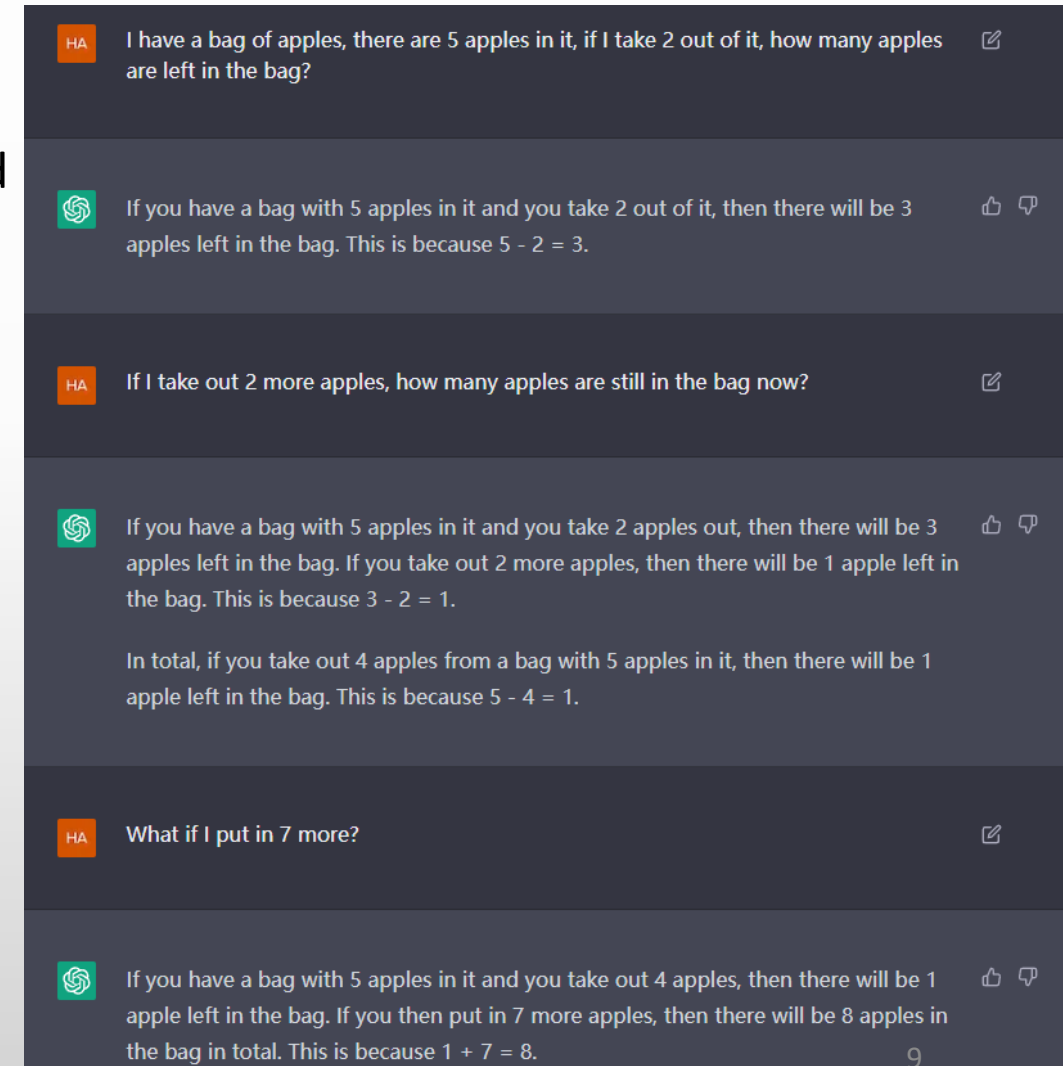
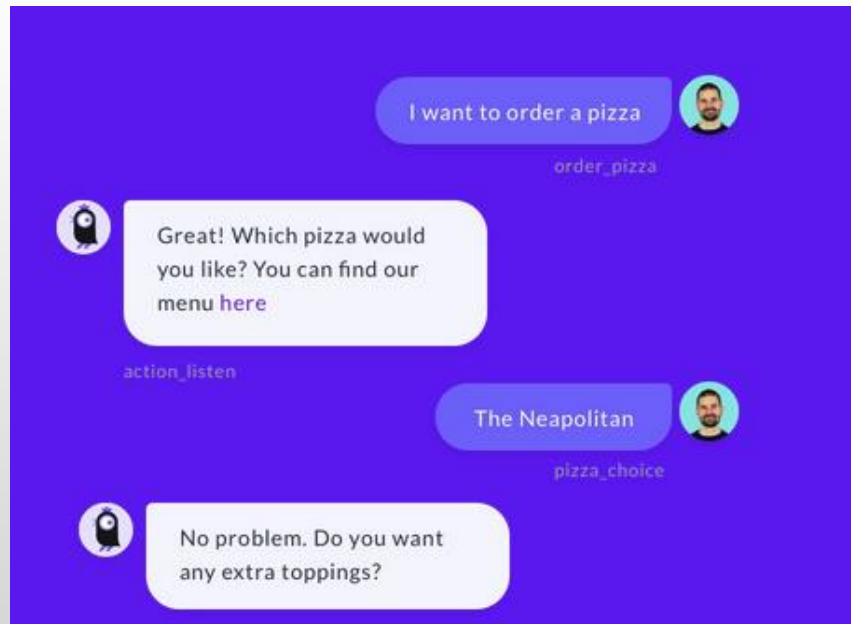
- NLP is a subfield in AI, where the goal is
  - To bridge the gap between how people communicate and what machines understand in order to perform useful tasks, e.g. making appointments, translation, question answering (QA), etc.





# Natural Language Processing (NLP)

- NLP is a subfield in AI, where the goal is
  - To bridge the gap between how people communicate and what machines understand in order to perform useful tasks, e.g. making appointments, translation, question answering (QA), etc.



# NLP Goals

Goal: intelligent processing of human language, not just string matching

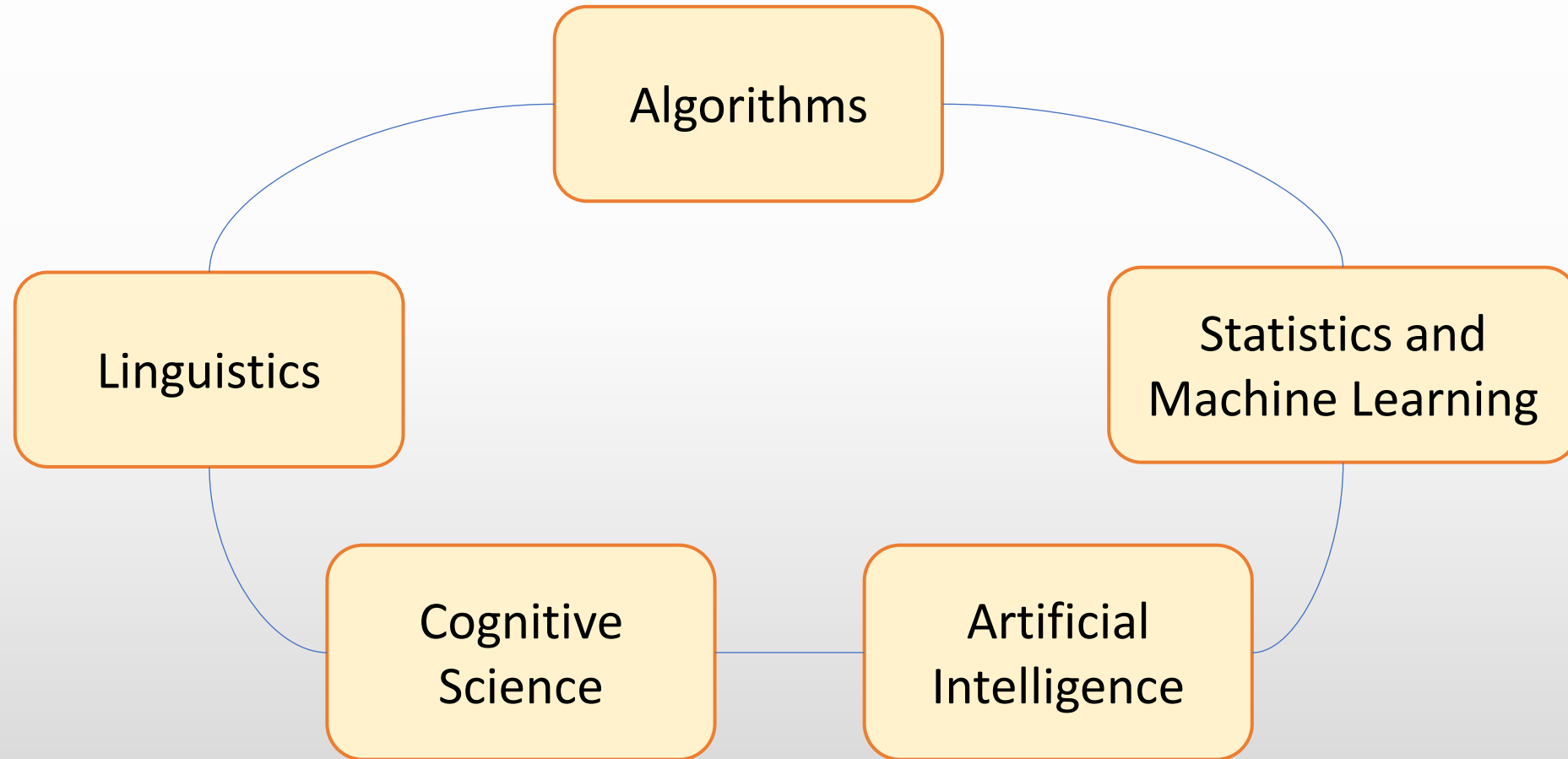


# NLP Goals

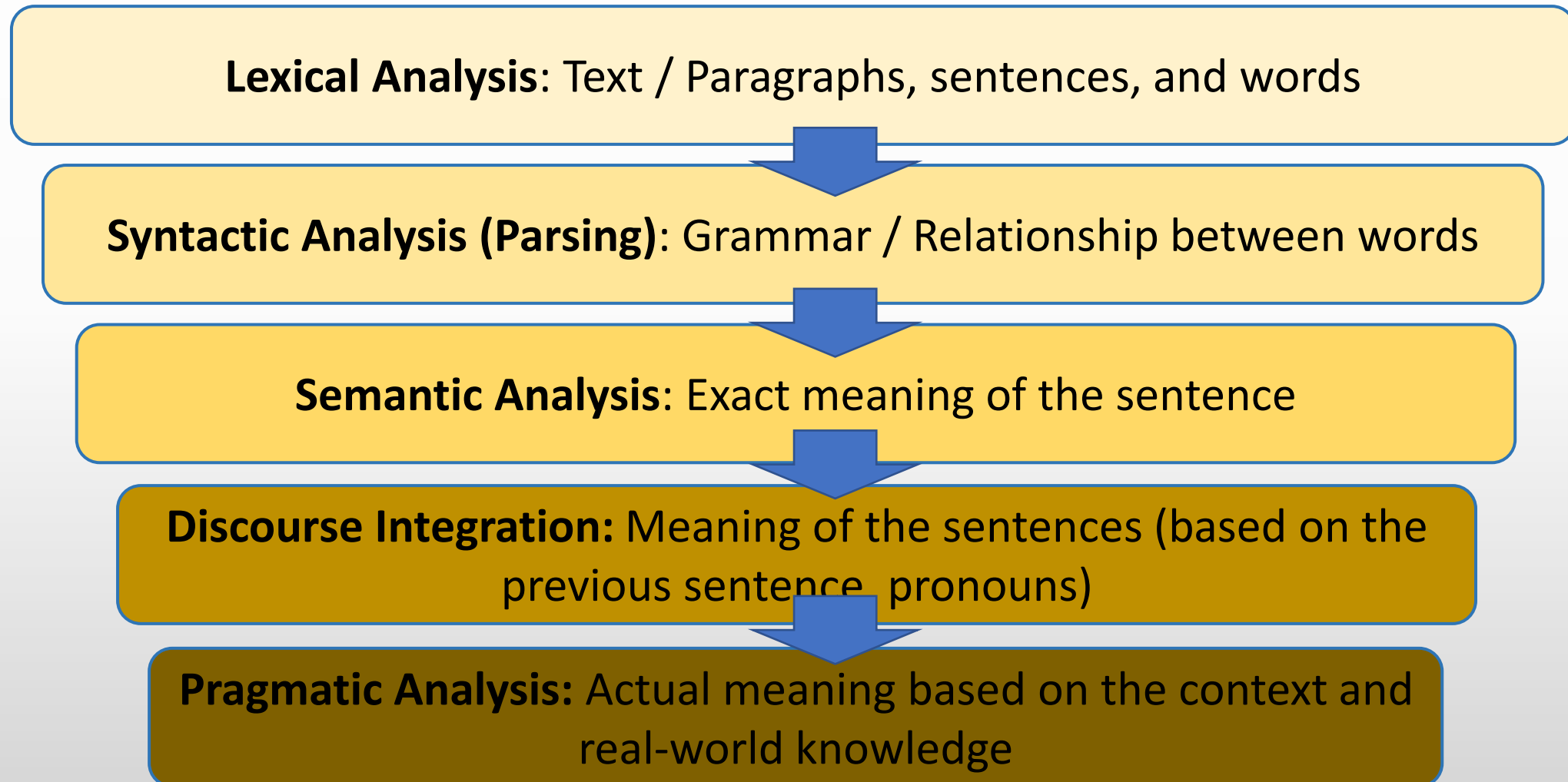
Goal: intelligent processing of human language, not just string matching



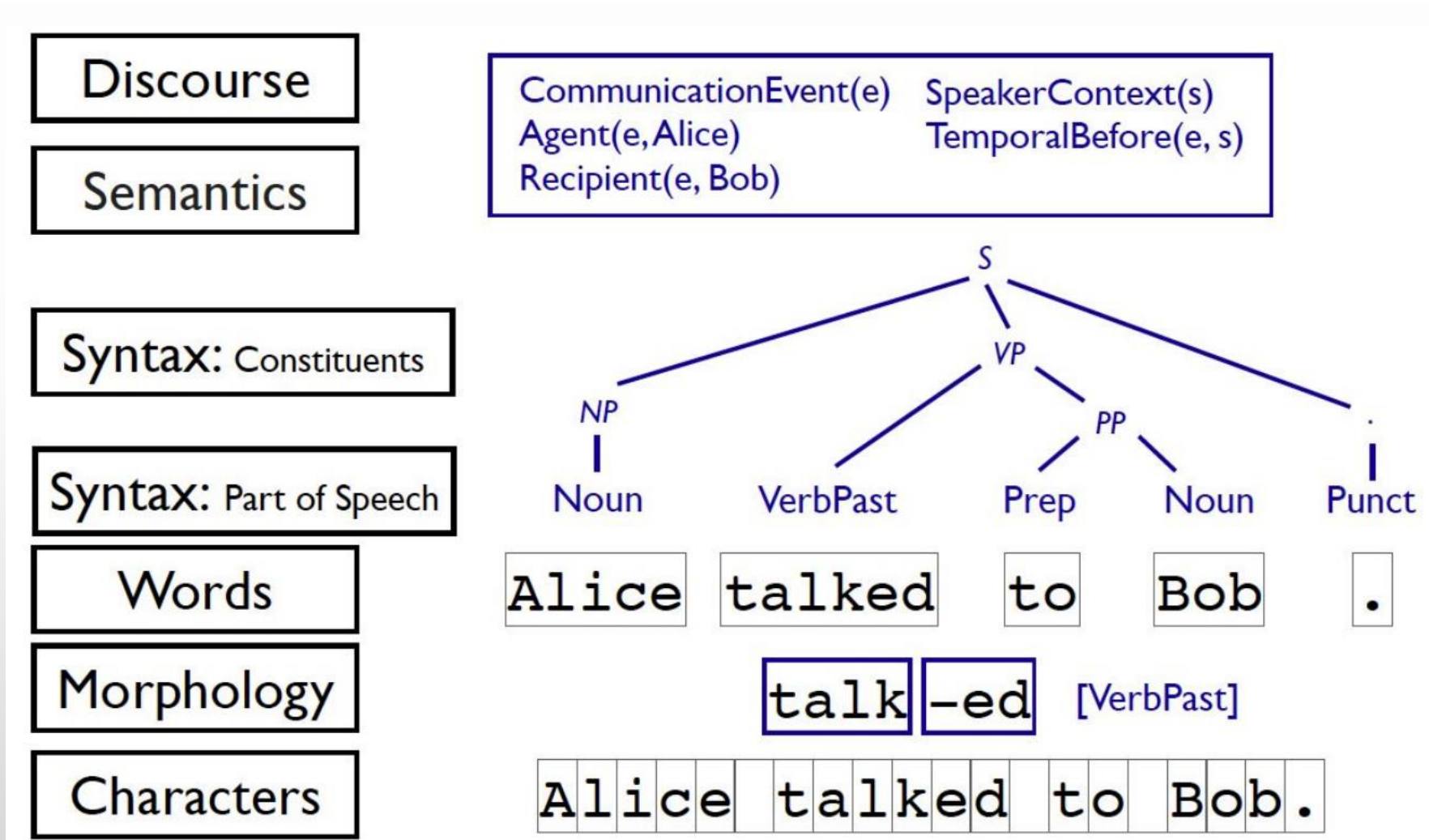
# NLP is interdisciplinary



# Level of understand in NLP




# Level of understand in NLP (cont.)





# NLP today and Technology

Dan Jurafsky



## Language Technology

making good progress

mostly solved

**Spam detection**

Let's go to Agra! ✓  
Buy V1AGRA ... ✗

**Part-of-speech (POS) tagging**

ADJ ADJ NOUN VERB ADV  
Colorless green ideas sleep furiously.

**Named entity recognition (NER)**

PERSON ORG LOC  
Einstein met with UN officials in Princeton

**Sentiment analysis**

Best roast chicken in San Francisco! 👍  
The waiter ignored us for 20 minutes. 👎

**Coreference resolution**

Carter told Mubarak he shouldn't run again.

**Word sense disambiguation (WSD)**

I need new batteries for my *mouse*.


**Parsing**

I can see Alcatraz from the window!

**Machine translation (MT)**

第13届上海国际电影节开幕... →  
The 13<sup>th</sup> Shanghai International Film Festival...

**Information extraction (IE)**

You're invited to our dinner party, Friday May 27 at 8:30  Party May 27 add

still really hard

**Question answering (QA)**

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?


**Paraphrase**

XYZ acquired ABC yesterday  
ABC has been taken over by XYZ

**Summarization**

The Dow Jones is up  
The S&P500 jumped  
Housing prices rose → Economy is good

**Dialog**

Where is Citizen Kane playing in SF?  
Castro Theatre at 7:30. Do you want a ticket? 

# NLP today: Machine Translation (MT)

The screenshot displays the Google Translate interface. At the top, there's a menu icon and the 'Google Translate' logo. Below this are three tabs: 'Text', 'Documents', and 'Websites'. The main interface is divided into two sections for language selection: 'DETECT LANGUAGE' and 'ENGLISH' on the left, and 'THAI', 'ENGLISH', and 'SPANISH' on the right. The 'ENGLISH' and 'THAI' options are selected. The left pane contains the English text: 'Global warming, caused by the increasing levels of greenhouse gases in the atmosphere, poses a significant danger to our planet. Rising temperatures can lead to more extreme weather events, such as heatwaves, droughts, and floods, which can have devastating effects on agriculture, human health, and infrastructure. Warmer oceans can cause coral reefs to die and lead to sea level rise, threatening coastal communities and low-lying areas. Additionally, as the polar ice caps melt, it can disrupt the delicate balance of marine ecosystems and lead to the extinction of many species. Overall, global warming is a serious threat to our environment and our way of life, and it is crucial that we take action to reduce our carbon emissions and slow the rate of warming.' The right pane shows the Thai translation: 'ภาวะโลกร้อน เกิดจากระดับก๊าซเรือนกระจกที่เพิ่มขึ้นในชั้นบรรยากาศ ก่อให้เกิดอันตรายอย่างมากต่อโลกของเรา อุณหภูมิที่สูงขึ้นสามารถนำไปสู่เหตุการณ์สภาพอากาศที่รุนแรงมากขึ้น เช่น คลื่นความร้อน ภัยแล้ง และน้ำท่วม ซึ่งอาจส่งผลกระทบร้ายแรงต่อการเกษตร สุขภาพของมนุษย์ และโครงสร้างพื้นฐาน มหาสมุทรที่อุ่นขึ้นอาจทำให้แนวปะการังตายและทำให้ระดับน้ำทะเลสูงขึ้น คุณภาพชุมชนชายฝั่งและพื้นที่ลุ่มต่ำ นอกจากนี้ เมื่อน้ำแข็งขั้วโลกละลาย มันสามารถทำลายสมดุลอันละเอียดอ่อนของระบบนิเวศทางทะเลและนำไปสู่การสูญพันธุ์ของสิ่งมีชีวิตหลายชนิด โดยรวมแล้ว ภาวะโลกร้อนเป็นภัยคุกคามร้ายแรงต่อสิ่งแวดล้อมและวิถีชีวิตของเรา และจำเป็นอย่างยิ่งที่เราต้องดำเนินการเพื่อลดการปล่อยก๊าซคาร์บอนและชะลออัตราการเกิดภาวะโลกร้อน'. Below the Thai text is a phonetic transcription in Thai script. At the bottom of the interface, there are icons for voice input/output, a character count '765 / 5,000', and a 'Show more' link.

Global warming, caused by the increasing levels of greenhouse gases in the atmosphere, poses a significant danger to our planet. Rising temperatures can lead to more extreme weather events, such as heatwaves, droughts, and floods, which can have devastating effects on agriculture, human health, and infrastructure. Warmer oceans can cause coral reefs to die and lead to sea level rise, threatening coastal communities and low-lying areas. Additionally, as the polar ice caps melt, it can disrupt the delicate balance of marine ecosystems and lead to the extinction of many species. Overall, global warming is a serious threat to our environment and our way of life, and it is crucial that we take action to reduce our carbon emissions and slow the rate of warming.

ภาวะโลกร้อน เกิดจากระดับก๊าซเรือนกระจกที่เพิ่มขึ้นในชั้นบรรยากาศ ก่อให้เกิดอันตรายอย่างมากต่อโลกของเรา อุณหภูมิที่สูงขึ้นสามารถนำไปสู่เหตุการณ์สภาพอากาศที่รุนแรงมากขึ้น เช่น คลื่นความร้อน ภัยแล้ง และน้ำท่วม ซึ่งอาจส่งผลกระทบร้ายแรงต่อการเกษตร สุขภาพของมนุษย์ และโครงสร้างพื้นฐาน มหาสมุทรที่อุ่นขึ้นอาจทำให้แนวปะการังตายและทำให้ระดับน้ำทะเลสูงขึ้น คุณภาพชุมชนชายฝั่งและพื้นที่ลุ่มต่ำ นอกจากนี้ เมื่อน้ำแข็งขั้วโลกละลาย มันสามารถทำลายสมดุลอันละเอียดอ่อนของระบบนิเวศทางทะเลและนำไปสู่การสูญพันธุ์ของสิ่งมีชีวิตหลายชนิด โดยรวมแล้ว ภาวะโลกร้อนเป็นภัยคุกคามร้ายแรงต่อสิ่งแวดล้อมและวิถีชีวิตของเรา และจำเป็นอย่างยิ่งที่เราต้องดำเนินการเพื่อลดการปล่อยก๊าซคาร์บอนและชะลออัตราการเกิดภาวะโลกร้อน

Phāwa lok fñn keid cāk radab kās reūxn krack thī pheīm khñn nī chñn brryākās kx hī keid xñntrāy xýāng māk lx lok kñxng reā xññphñmī thī sñng khñn sāmārth nā pī sññ fietukārñ sñphāph xākās thī runræng māk khñn chñn khññ khwām fñn phay læng læa kñ thwm

Show more

[Send feedback](#)



# NLP today: Question Answering (QA)



IBM Watson wowed the tech industry and a corner of U.S. pop culture with its 2011 win against two of Jeopardy's greatest champions. Here's how IBM pulled it off and a look at what Watson's real career is going to be.

<https://www.techrepublic.com/article/ibm-watson-the-inside-story-of-how-the-jeopardy-winning-supercomputer-was-born-and-what-it-wants-to-do-next/>

HA

I have a bag of apples, there are 5 apples in it, if I take 2 out of it, how many apples are left in the bag?

ChatGPT

If you have a bag with 5 apples in it and you take 2 out of it, then there will be 3 apples left in the bag. This is because  $5 - 2 = 3$ .

HA

If I take out 2 more apples, how many apples are still in the bag now?

If you have a bag with 5 apples in it and you take 2 apples out, then there will be 3 apples left in the bag. If you take out 2 more apples, then there will be 1 apple left in the bag. This is because  $3 - 2 = 1$ .

In total, if you take out 4 apples from a bag with 5 apples in it, then there will be 1 apple left in the bag. This is because  $5 - 4 = 1$ .







HA

What if I put in 7 more?

If you have a bag with 5 apples in it and you take out 4 apples, then there will be 1 apple left in the bag. If you then put in 7 more apples, then there will be 8 apples in the bag in total. This is because  $1 + 7 = 8$ .


17

# NLP today: Search & Summarization

 chatgpt     

[All](#) [News](#) [Images](#) [Videos](#) [Books](#) [More](#) [Tools](#)

About 234,000,000 results (0.34 seconds)



<https://openai.com> > [blog](#) > chatgpt 


## ChatGPT: Optimizing Language Models for Dialogue - OpenAI

Nov 30, 2565 BE — We've trained a model called **ChatGPT** which interacts in a conversational way. The dialogue format makes it possible for **ChatGPT** to answer ...


[InstructGPT](#) · [OpenAI - OpenAI](#) · [Proximal Policy Optimization](#) · [Moderation API](#)

You've visited this page many times. Last visit: 1/17/23


 **Top stories** 

 TNN


Microsoft เตรียมขยายการเข้าถึงโปรแกรม ChatGPT ให้ลูกค้าทั่วไปใช้งาน




1 hour ago

 Siam Blockchain


สาวคนหนึ่งให้ ChatGPT ทำนายราคา Bitcoin และเธอก็ได้คำตอบที่เหนือความคาด...




3 days ago


 Bangkok Post

AI, do my homework! How ChatGPT pitted teachers against tech



 Siam Blockchain

สื่อข่าวชื่อดังถาม ChatGPT ว่าแนวโน้มราคา Bitcoin ในปี 2030 จะเป็นอย่างไรและได้



## ChatGPT

Computer program



ChatGPT is a chatbot launched by OpenAI in November 2022. It is built on top of OpenAI's GPT-3 family of large language models, and is fine-tuned with both supervised and reinforcement learning techniques. [Wikipedia](#)

**Initial release:** November 30, 2022; 28 days ago

**Developer(s):** [OpenAI](#)

**License:** Proprietary

People also search for [View 10+ more](#)

# NLP today: Information Extraction

Abstract clinical report into a database



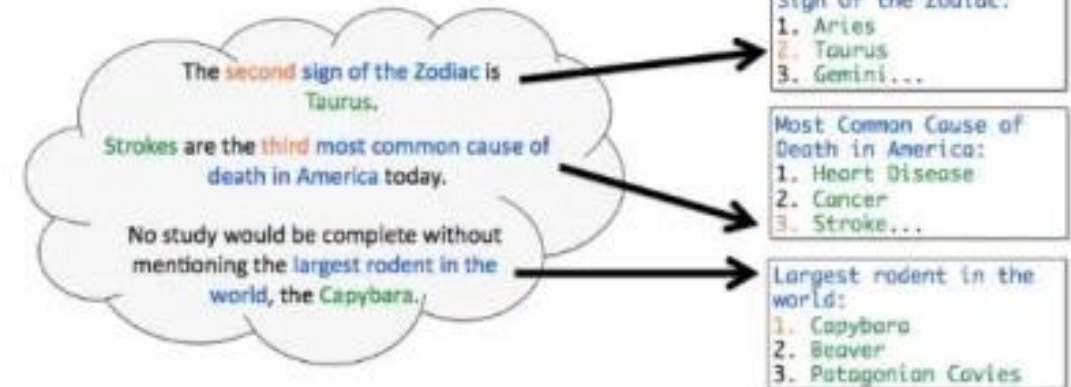
ID	AGE	RACE	STUDY	PROC	BIRTHS	MA AGE	ASSESS	DENSITY	FINDING	FINDING T
9527	78	2	6/12/06	BIDXL-L	0		P	3	CALCS	N
32875	56	1	7/11/06	BIDXB-B	0		N	3		
2247	72	1	4/12/06	BIDXL-R	0		N	3		
45521	61	1	3/30/06	BIDXB-B	0		B	3	CALCS	S
48987	41	1	4/5/06	BIDXB-B	0		P	3	CALCS	N
4179	67	1	5/12/06	BIDXB-B	0		P	2	CALCS	N
26300	59	1	3/31/06	BIDXL-L	0		N	3		
67960	64	1	4/7/06	BIDXL-R	0		P	3	MASS	O
43283	61 W		7/21/06	BIDXB-B	0		B	3		
43319	51	1	4/7/06	BIDXB-B	0		N	3		

## What is IE?

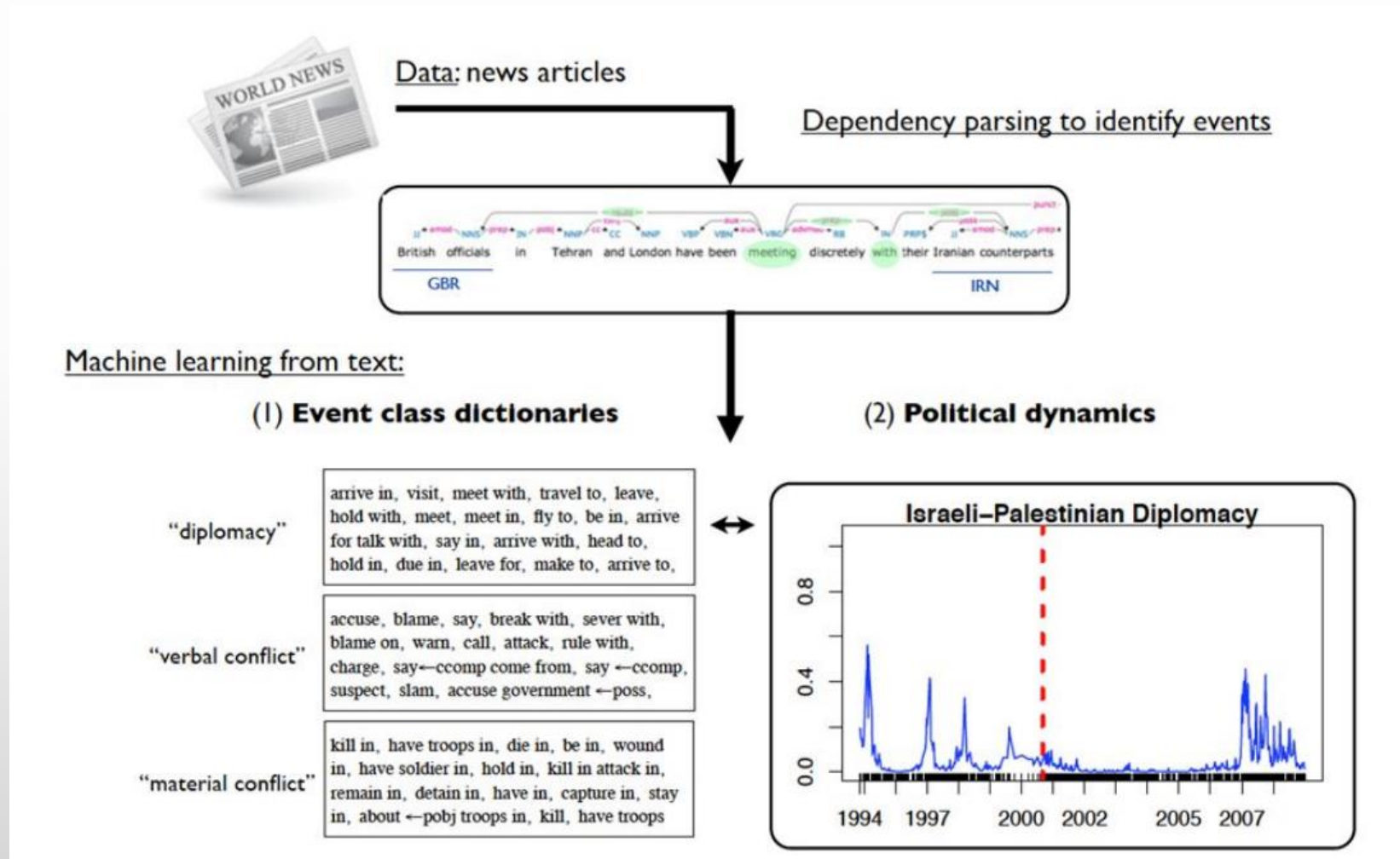
Unstructured  
Web Text



Structured  
Sequences



# NLP today: Trend analysis



# NLP today: Trend analysis (cont.)

## Hathaway Phenomenon



A couple weeks ago, Huffington Post blogger Dan Mirvish noted a funny trend: when Anne Hathaway was in the news, Warren Buffett's Berkshire Hathaway's shares went up. He pointed to [six dates going back to 2008](#) to show the correlation. Mirvish then suggested a mechanism to explain the trend: "automated, robotic trading programming are picking up the same chatter on the Internet about 'Hathaway' as the IMDb's StarMeter, and they're applying it to the stock market." Ref: Prof. Regina Barzilay, NLP @MIT



# NLP today: Trend analysis (correlation)

## Hathaway Phenomenon



BERKSHIRE  
HATHAWAY



A couple weeks ago, Huffington Post blogger Dan Mirvish noted a funny trend: when Anne Hathaway was in the news, Warren Buffett's Berkshire Hathaway's shares went up. He pointed to [six dates going back to 2008](#) to show the correlation. Mirvish then suggested a mechanism to explain the trend: "automated, robotic trading programming are picking up the same chatter on the Internet about 'Hathaway' as the IMDb's StarMeter, and they're applying it to the stock market." Ref: Prof. Regina Barzilay, NLP @MIT

### BERKSHIRE HATHAWAY INC.

3555 Farnam Street  
Omaha, NE 68131  
Official Home Page

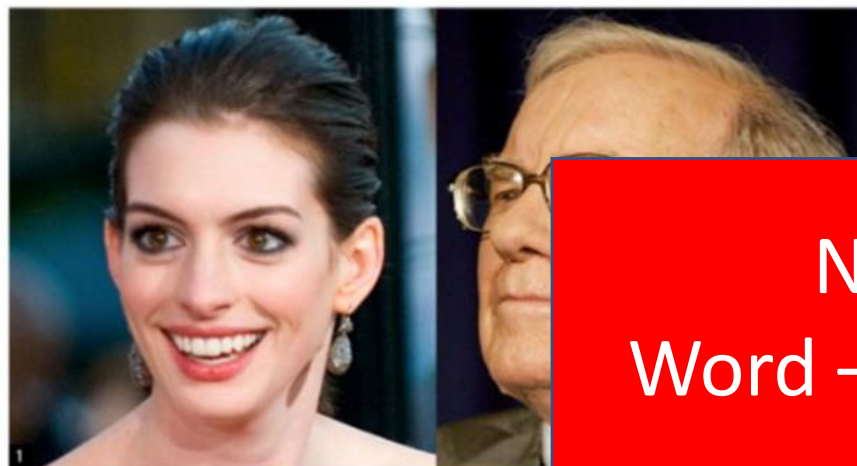
- [A Message from Warren E. Buffett](#)
- [Annual & Interim Reports](#)  
Updated November 5, 2022
- [Link to SEC Filings](#)
- [Special Letters from Warren & Charlie RE: Past, Present and Future](#)
- [Links to Berkshire Subsidiary Companies](#)
- [Corporate Governance](#)
- [Sustainability](#)
- [Information Regarding Warren E. Buffett's Annual Contributions to Five Foundations](#)
- [Berkshire Activewear](#)
- [News Releases from Berkshire Hathaway and from Warren Buffett](#)  
Updated December 19, 2022
- [Annual Meeting Information](#)  
Updated October 20, 2022
- [Berkshire Hathaway Energy Investor Presentations](#)
- [Warren Buffett's Letters to Berkshire Shareholders](#)  
Updated February 26, 2022
- [Charlie Munger's Letters to Wesco Shareholders](#)
- [Celebrating 50 Years of a Profitable Partnership](#)  
(A commemorative book first sold at the 2015 Annual Meeting and now for sale on eBay.)
- [Common Stock Information](#)
- [Facts Regarding Berkshire's 2021 Investments in Activision Common Stock](#)

# NLP today: Trend analysis (correlation)

## Hathaway Phenomenon



BERKSHIRE  
HATHAWAY



NLP is difficult.  
Word –level ambiguity !!!

A couple weeks ago, Huffington Post blogger Dan Mirvish noted a funny trend: when Anne Hathaway was in the news, Warren Buffett's Berkshire Hathaway's shares went up. He pointed to [six dates going back to 2008](#) to show the correlation. Mirvish then suggested a mechanism to explain the trend: "automated, robotic trading programming are picking up the same chatter on the Internet about 'Hathaway' as the IMDb's StarMeter, and they're applying it to the stock market." Ref: Prof. Regina Barzilay, NLP @MIT

BERKSHIRE HATHAWAY INC.  
3555 Farnam Street  
Omaha, NE 68131  
Official Home Page

Chairman and CEO: Warren E. Buffett

• [Corporate Governance](#)

• [Sustainability](#)

• [Information Regarding Warren E. Buffett's Annual Contributions to Five Foundations](#)

• [Berkshire Activewear](#)

• [News Releases from Berkshire Hathaway and from Warren Buffett](#)  
Updated December 19, 2022

• [Annual Meeting Information](#)  
Updated October 20, 2022

• [Berkshire Hathaway Energy Investor Presentations](#)

• [Warren Buffett's Letters to Berkshire Shareholders](#)  
Updated February 26, 2022

• [Charlie Munger's Letters to Wesco Shareholders](#)

• [Celebrating 50 Years of a Profitable Partnership](#)  
(A commemorative book first sold at the 2015 Annual Meeting and now for sale on eBay.)

• [Common Stock Information](#)

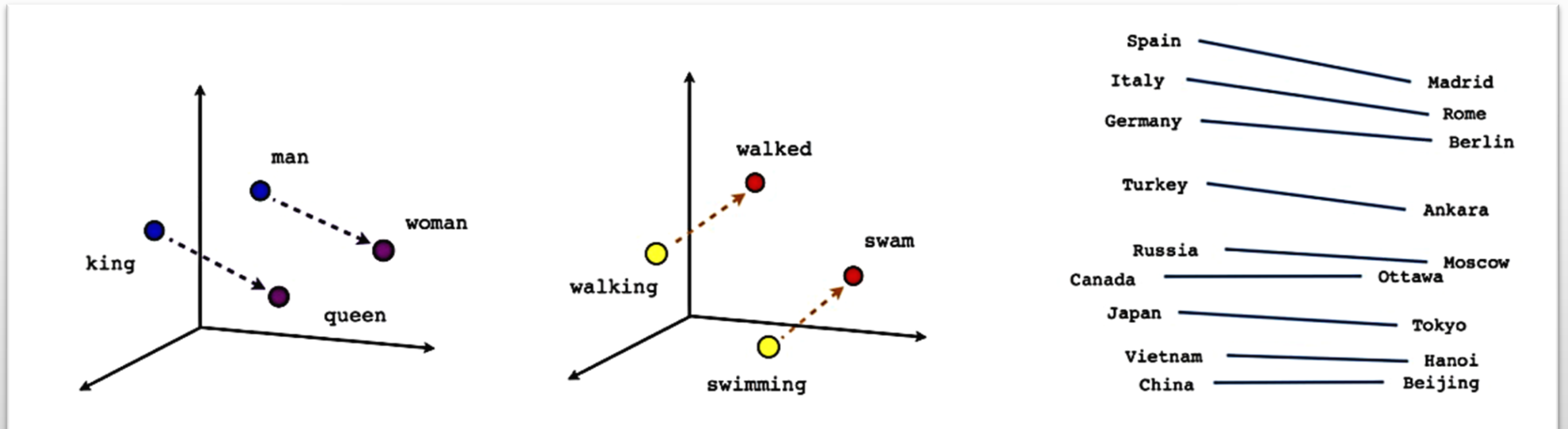
• [Facts Regarding Berkshire's 2021 Investments in Activision Common Stock](#)

Why it is difficult ?

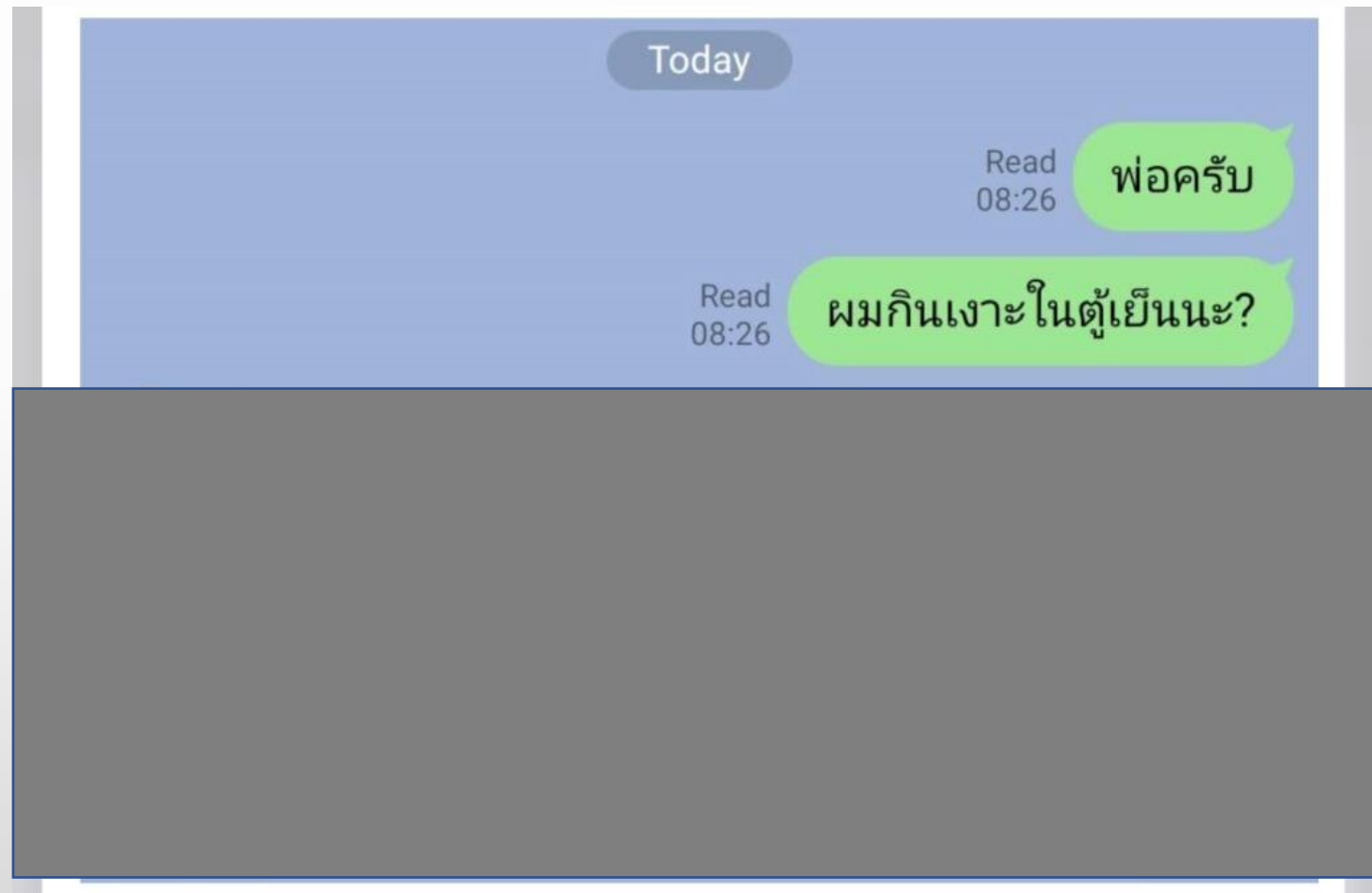


# What NLP is difficult?

- The complexity in expressing, acquiring, and utilizing knowledge related to language, context, the world, and visual information.



# What NLP is difficult? (cont.)



# What NLP is difficult? (cont.)



<https://www.sti.chula.ac.th/knowledge/2820/>

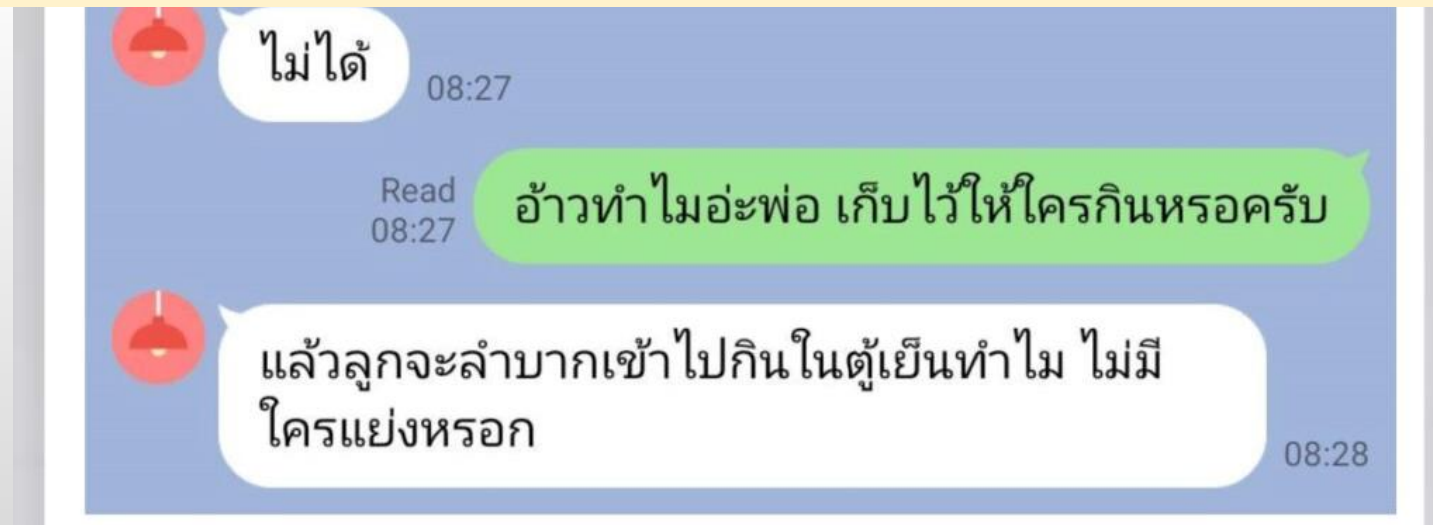
# What NLP is difficult? (cont.)



<https://www.sti.chula.ac.th/knowledge/2820/>

# What NLP is difficult? (cont.)

- Human language are ambiguous (unlike programming and other formal language), so some parts can be ignored.
- Human language are interpretation depending on real world, common sense, and contextual knowledge.



# What NLP is difficult? (cont.)

## Word segmentation (No word delimiters)

- ฉัน | กำลัง | จะ | ไป | เรียน | วิชา | ประมวล | ผลภาษา | ธรรมชาติ | ที่ | สถาบัน | วิทยาการ | หุ่น | ยนต์ | ภาค | สนาม
- ฉัน | กำลัง | จะ | ไป | เรียน | วิชา | ประมวลผลภาษา | ธรรมชาติ | ที่ | สถาบัน | วิทยาการ | หุ่นยนต์ | ภาคสนาม
- ฉัน | กำลังจะ | ไป | เรียน | วิชา | ประมวลผลภาษาธรรมชาติ | ที่ | สถาบันวิทยาการหุ่นยนต์ภาคสนาม

# What NLP is difficult? (cont.)

## Word segmentation (No word delimiters)

- ฉัน | กำ | ลัง | จะ | ไป | เรียน | วิชา | ประมวล | ผลภาษา | ธรรมชาติ | ที่ | สถาบัน | วิทยาการ | หุ่น | ยนต์ | ภาค | สนาม
- ฉัน | **กำ** | **ลัง** | จะ | ไป | เรียน | วิชา | **ประมวลผลภาษา**
- ฉัน | **กำ** | **ลัง** | **จะ** | ไป | เรียน | วิชา | **ประมวลผลภาษา**

## Sentence segmentation (No sentence boundary markers)

<https://th.wikipedia.org/wiki/%E0%B9%81%E0%B8%8A%E0%B8%97%E0%B8%88%E0%B8%B5%E0%B8%9E%E0%B8%B5%E0%B8%97%E0%B8%B5>

แชทจีพีที (อังกฤษ: ChatGPT) เป็น**แชทบอทปัญญาประดิษฐ์**ต้นแบบที่พัฒนาโดย**โอเพนเอไอ** ซึ่งเชี่ยวชาญด้านการสนทนา แชทบอทเป็น**โมเดลภาษาขนาดใหญ่**ที่ได้รับการ**ปรับแต่งอย่างละเอียด**ทั้งการควบคุมและ**เทคนิคการเรียนรู้แบบเสริมแรง** มันอยู่บนพื้นฐานของรุ่น GPT-3.5 ของโอเพนเอไอซึ่งเป็นรุ่นปรับปรุงของ GPT-3

แชทจีพีทีเปิดตัวในเดือนพฤศจิกายน ค.ศ. 2022 และได้รับความสนใจจากรายละเอียดการตอบสนองและคำตอบที่ชัดเจน แม้ว่าความถูกต้องตามข้อเท็จจริงนั้นจะถูกวิพากษ์วิจารณ์ก็ตาม

### การตอบรับ [แก้]

ในเดือนธันวาคม ค.ศ. 2022 มีความเห็นบางส่วนเช่น **พอล ครุกแมน** นักเศรษฐศาสตร์ เขียนว่า แชทจีพีที จะส่งผลกระทบต่อความต้องการของ**ผู้ปฏิบัติงานด้านความรู้**<sup>[1]</sup> ใน **เดอะเวิร์จ** เจมส์ วินเซนต์ มองเห็นความสำเร็จอย่างแพร่หลายของแชทจีพีทีเป็นหลักฐานว่าปัญญาประดิษฐ์กลายเป็นกระแสหลักไปแล้ว<sup>[2]</sup> ใน **ดิแอตแลนติก** สตีเฟน มาร์เช่ ตั้งข้อสังเกตว่ามีผลกระทบต่อสถาบันการศึกษาและโดยเฉพาะอย่างยิ่ง**เรียงความการสมัครงาน**ยังไม่เป็นที่เข้าใจ<sup>[3]</sup> แดเนียล เฮอร์แมน ครูและนักเขียนของโรงเรียนมัธยมในแคลิฟอร์เนียเขียนว่า แชทจีพีทีจะนำเข้าสู่ "จุดจบของภาษาอังกฤษระดับมัธยมปลาย"<sup>[4]</sup>

# What NLP is difficult? (cont.)

- Syntax ambiguity

อาจารย์ป้อออกข้อสอบวิชา **NLP** ให้ยาก เพราะ ..... เชื่อว่า นศ ทุกคนจะทำ..... ได้



# What NLP is difficult? (cont.)

- Syntax ambiguity

อาจารย์ป้อออกข้อสอบวิชา **NLP** ให้ยาก เพราะ ..... เชื่อว่า นศ ทุกคนจะทำ..... ได้

อาจารย์ป้อออกข้อสอบวิชา **NLP** ให้ยาก เพราะ ..... เชื่อว่า นศ ทุกคนจะทำ..... ได้



# What NLP is difficult? (cont.)

- Syntax ambiguity

อาจารย์ป้อออกข้อสอบวิชา **NLP** ให้ยาก เพราะ ..... เชื่อว่า นศ ทุกคนจะทำ..... ได้

อาจารย์ป้อออกข้อสอบวิชา **NLP** ให้ยาก เพราะ ..... เชื่อว่า นศ ทุกคนจะทำ..... ได้



อาจารย์ป้อออกข้อสอบวิชา **NLP** ให้ยาก เพราะ เชื่อว่า นศ ทุกคนจะทำได้

# What NLP is difficult? (cont.)

- Slang

นร: ทำไมอาจารย์ออกข้อสอบวิชา **NLP** ยากจังครับ

อาจารย์ปอ: เพราะอาจารย์ **เลือดกรุ๊ปบี**

A: "I can't decide whether to take the job offer or not, what should I do?"

B: "Have you considered all the pros and cons of the job?"

A: "Yes, I have but I am still unsure"

B: "Well, why don't you **sleep on it**? It's a big decision, and it's important to take your time and make sure you're comfortable with your choice."


A: "That's a good idea. I'll think about it tonight and let you know tomorrow."


# Processing Steps of NLP


# Processing steps of NLP


- NLP is to learning the meaning.


**Tweets**

 **Jeremiah Owyang** @jowyang 5s  
Who's my Community Manager hero? @MJensen of @Yammer was listening and responded, I called him to thank! #CMAD tx @thismuchweknow  
Expand Reply Retweet Favorite More

 **Aaron Lee** @AskAaronLee 6s  
I love my quiet mornings. You?  
Expand Reply Delete Favorite More

 **Jay Rosen** @jayrosen\_nyu 23s  
The simplest way to say what I think most —not all —are missing in @ezraklein's new venture is... it's a UI play. (User interface for news.)  
Expand Reply Retweet Favorite More

 **Lisa Petrilli** @LisaPetrilli 57s  
Breaking Up Google goo.gl/M5uOFX via @Tenacity5Media  
View summary Reply Retweet Favorite Pocket More

 **Mashable** @mashable 1m  
Australian Open Champ Delivers Hilarious Victory Speech on.mash.to/Mn7rzc  
View summary Reply Retweet Favorite Pocket More

Unstructured data



Comments	Good	Like	Hate	Meaning
Tweet1	10	5	2	😊
Tweet2	5	3	8	😞
Tweet3	6	4	3	😊

Structured data

# Processing steps of NLP (cont.)

N: Proper Noun  
V: Main verb  
P: Preposition  
PER: Person  
LOC: Location  
O: Other

## Tokenization

Input: Mr. Paisit goes to Japan  
Output: [Mr. Paisit, goes, to, Japan]

## Part of Speech tagging

Input: [Mr. Paisit, goes, to, Japan]  
Output: [(Mr. Paisit, N), (goes, V), (to, P), (Japan, N)]

## NER

Input: [(Mr. Paisit, N), (goes, V), (to, C), (Japan, N)]  
Output: [(Mr. Paisit, N, PER), (goes, V, O), (to, P, O), (Japan, N, LOC)]

## Application - e.g. Word Cloud




# NLP Approaches


- Symbolic approach
  - The symbolic approach involves inputting all necessary information into a computer.
  - This approach led to the prominence of knowledge-based methods, which rely on human experts to input knowledge into the computer.
  - Disadvantage: It is required substantial human effort.

Goal: It aims to place “the” (determiner).

Scientists in United States have found way of turning lazy monkeys into workaholics using gene therapy. Usually monkeys work hard only when they know reward is coming, but animals given this treatment did their best all time. Researchers at National Institute of Mental Health near Washington DC, led by Dr Barry Richmond, have now developed genetic treatment which changes their work ethic markedly. "Monkeys under influence of treatment don't procrastinate," Dr Richmond says. Treatment consists of anti-sense DNA - mirror image of piece of one of our genes - and basically prevents that gene from working. But for rest of us, day when such treatments fall into hands of our bosses may be one we would prefer to put off.

 **DEFINITE ARTICLE: THE**

The **definite article** is the word "**the**". It is used before a noun to specify it as something previously considered.

When to Use "THE":

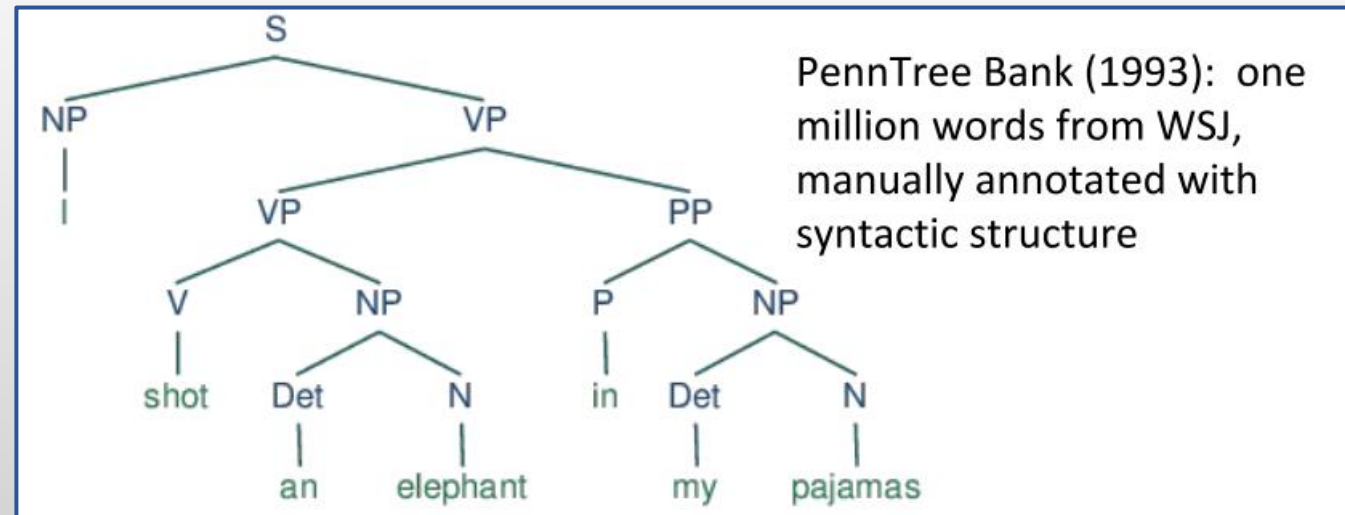
- Use "the" after you have talked about it the first time
  - This is a dog. The dog is brown.
- Use "the" when the person you are talking to already knows what you are talking about.
  - I am going to the bank.
  - There is someone knocking at the door. It must be the computer repairman.
- Use "the" when you talk about general places in nature like the lake, the mountains, the ocean, the beach, etc.
  - Let's go to the beach.
  - I love to camp in the mountains.
- Use "the" when you are talking about something that there is only one of and everyone knows about it. The sky, the earth, the stars (as a group).
  - How many people are there on the Earth?
  - The sky is so beautiful today.

www.englishstudyonline.org

39

# NLP Approaches (cont.)

- Statistical or Machine learning approach
  - utilizes language samples to understand the properties of language.
  - In the 1980s, there was a shift towards using probabilistic methods in NLP, which was inspired by information theory.
  - However, this approach has the drawback that it requires the manual creation of features.

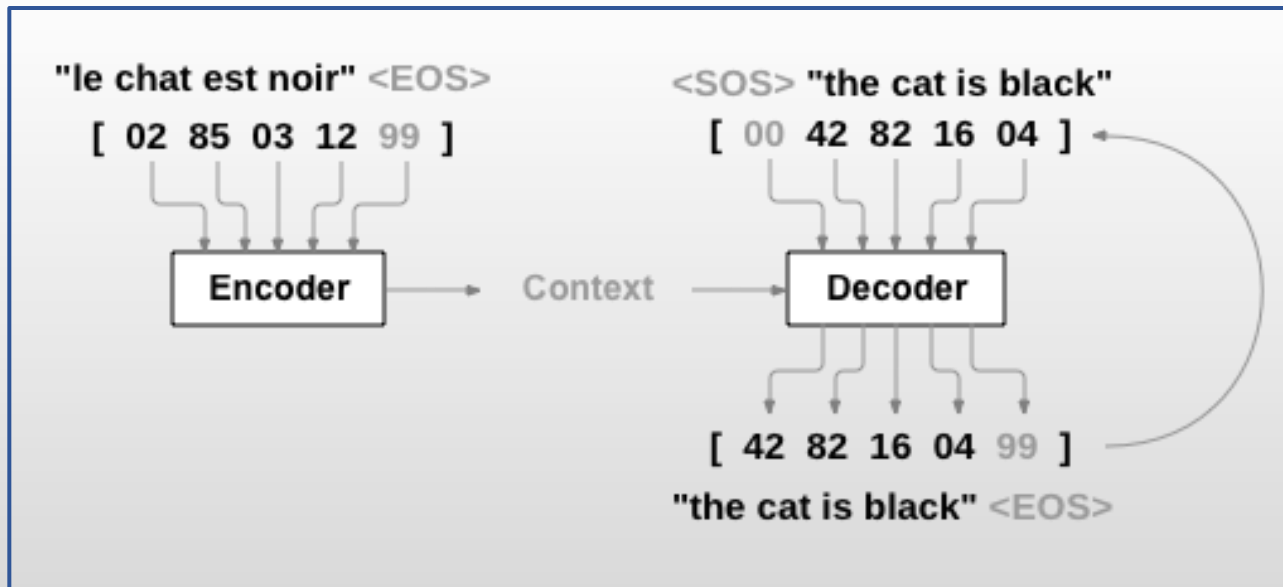




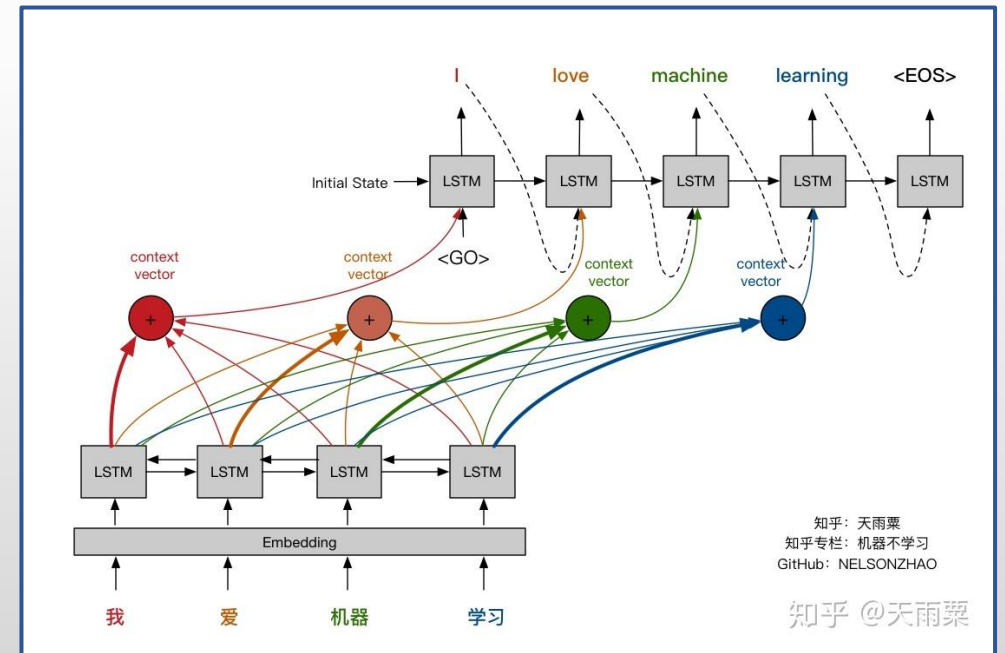
# NLP Approaches (cont.)

- Deep learning approach

The approach in question is a neural network that incorporates feature engineering, and it has been increasingly popular and successful since 2010.



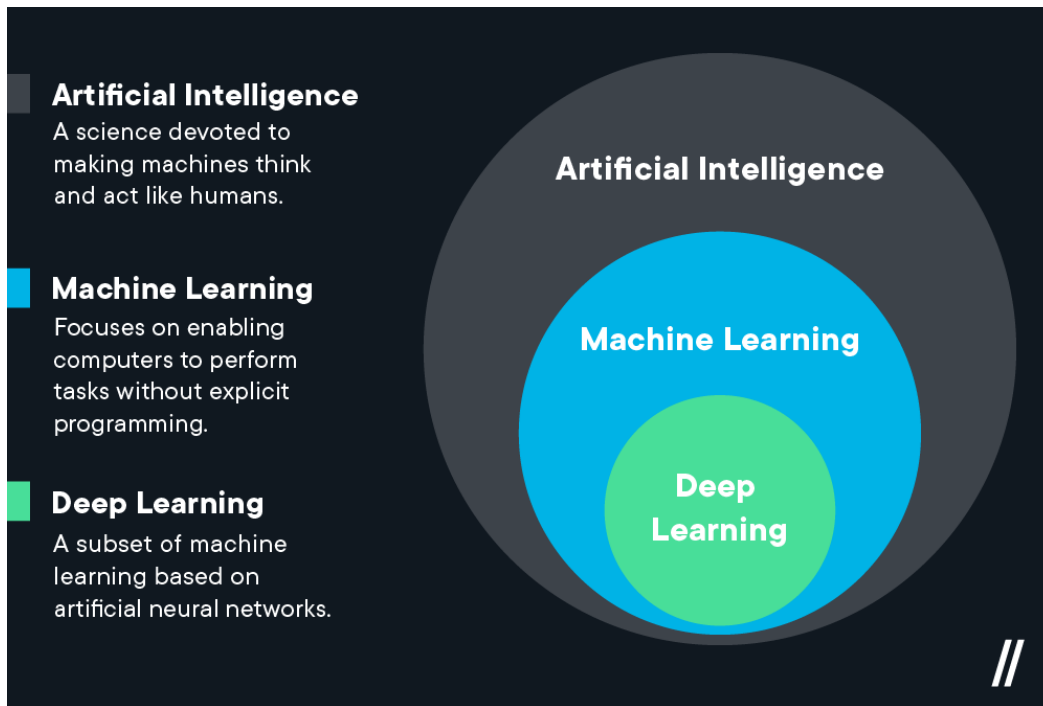
Sequence to sequence model



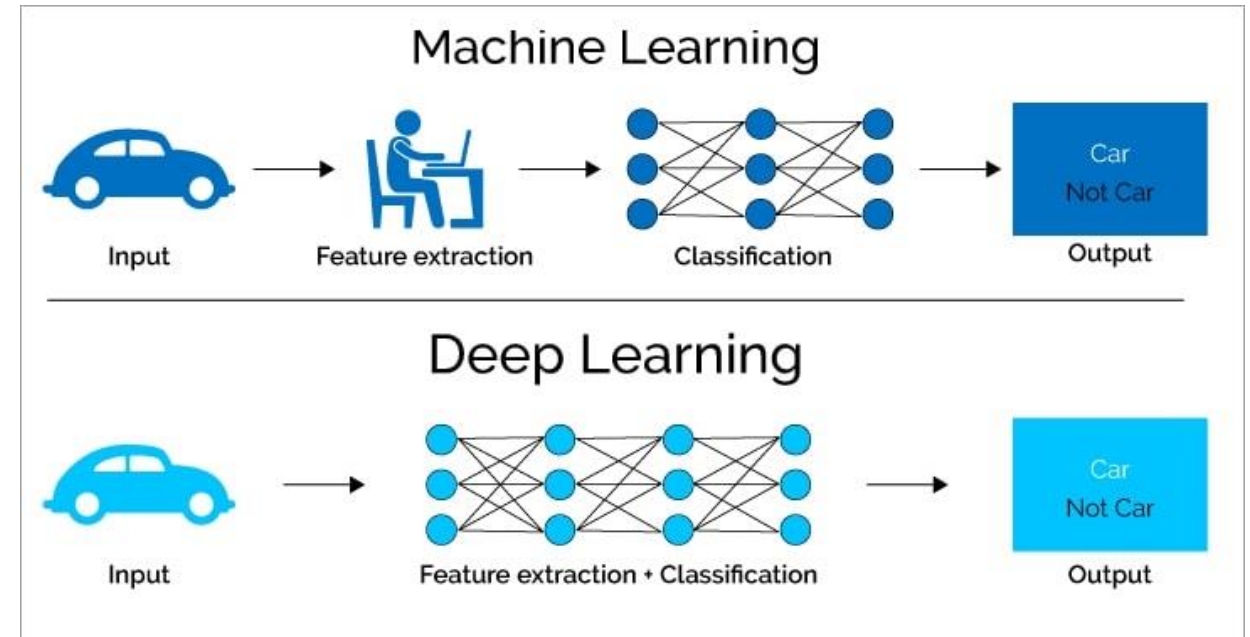
Attention mechanism in sequence to sequence model

# Deep Learning in NLP

# What is Deep Learning?

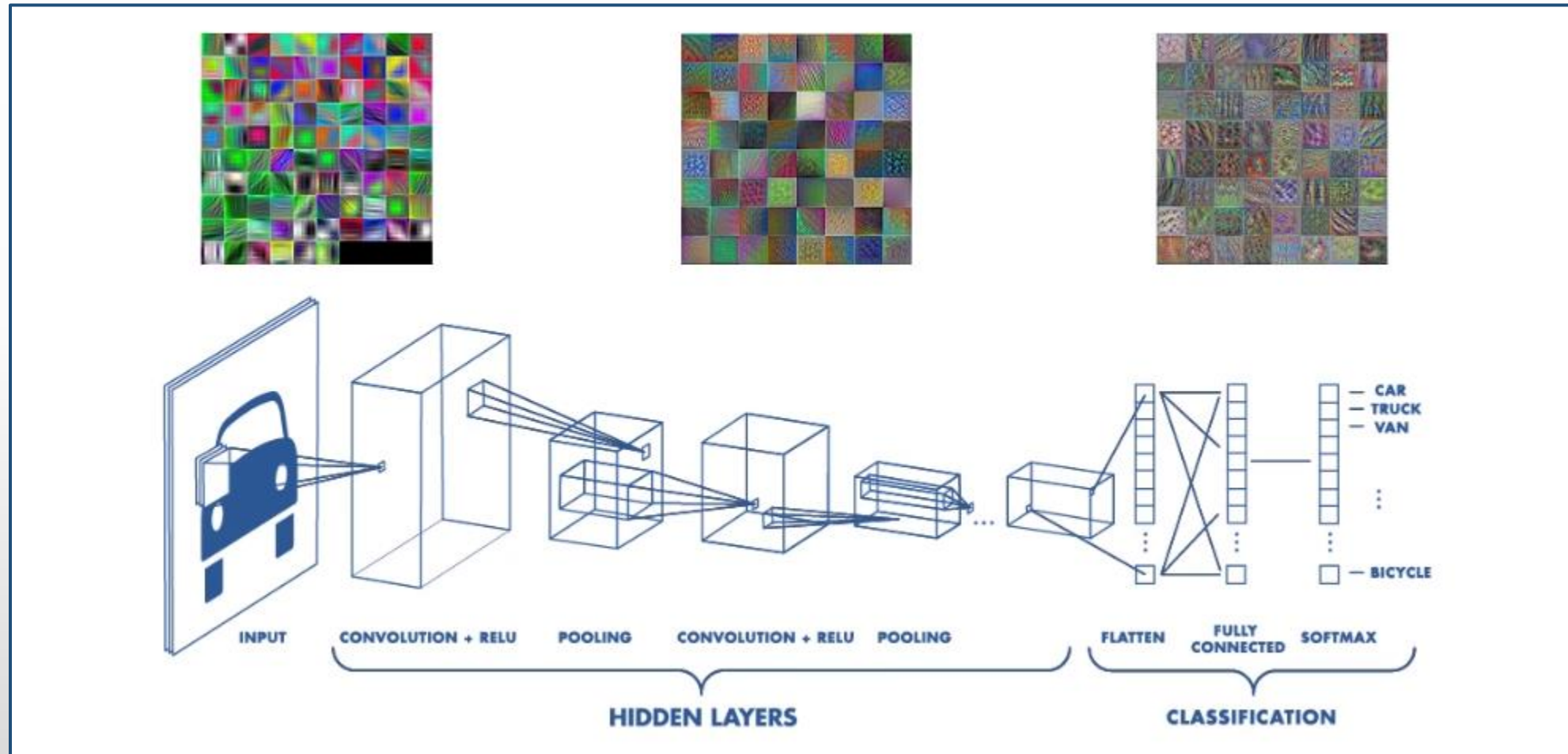


<https://flatironschool.com/blog/deep-learning-vs-machine-learning/>



<https://levity.ai/blog/difference-machine-learning-deep-learning>

# Deep Learning Basics



<https://python.plainenglish.io/convolution-neural-network-cnn-in-deep-learning-77f5ab457166>

# NLP + Deep Learning = Deep NLP

- The current methods used in NLP are based on advanced deep learning techniques.
- These methods have shown exceptional results in a wide range of NLP tasks.
- They typically do not need manual manipulation of linguistic features to achieve good performance.

<https://www.youtube.com/watch?v=V8qrVleGY5U>

## [PDF] Deep learning for NLP

R Socher, Y Bengio, C Manning - Tutorial at Association of Computational ..., 2012 - Citeseer  
... #4 Using a **deep** architecture ... #6 **Deep NLP** Learning models ... #6 **Deep NLP**

☆ Save Cite Cited by 48 Related articles All 4 versions

## How transfer learning impacts linguistic knowledge in **deep NLP** models?

N Durrani, H Sajjad, F Dalvi - arXiv preprint arXiv:2105.15179, 2021 - arxiv.org

... a predominant theme in **NLP** recently. Several researchers have shown that **deep NLP** models learn ... We investigate how fine-tuning towards downstream **NLP** tasks impacts the learned ...

☆ Save Cite Cited by 27 Related articles All 6 versions

## An overview of shallow and **deep natural language processing** for ontology learning

A Zouag - Ontology learning and knowledge discovery using the ..., 2011 - igi-global.com

... **natural language processing (NLP)** techniques ranging from lexical acquisition to shallow and **deep** ... when they come as a complement to the **NLP** process. There are many reasons why ...

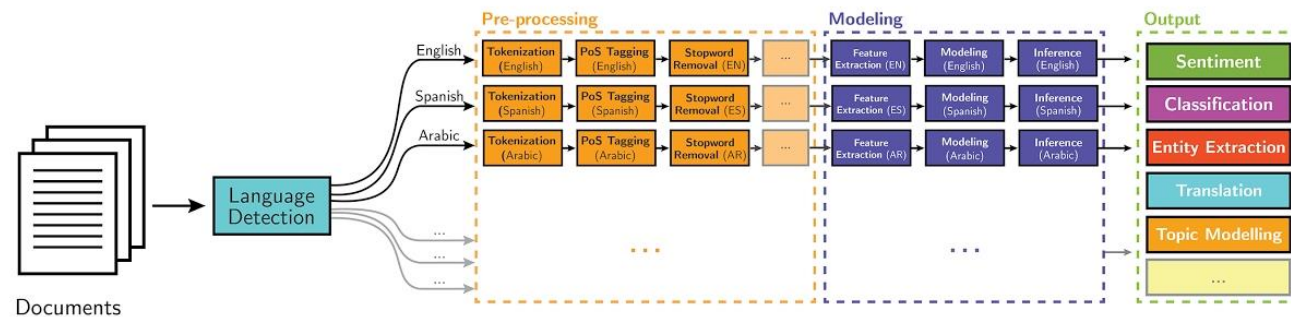
☆ Save Cite Cited by 34 Related articles

## Neuron-level interpretation of **deep nlp** models: A survey

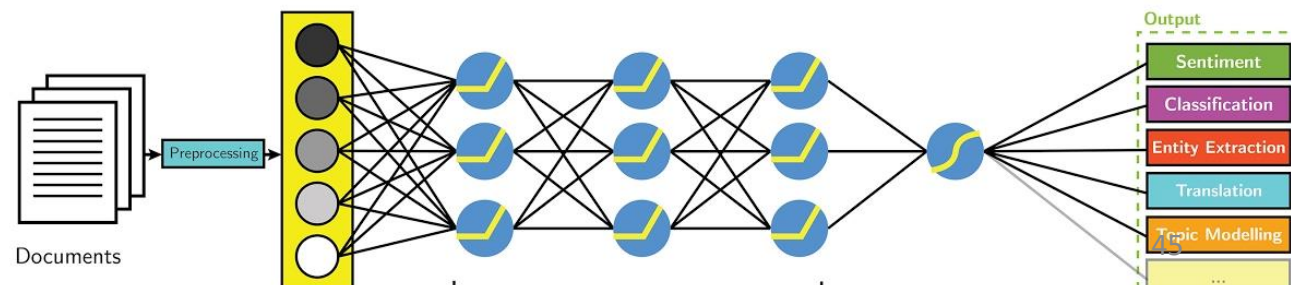
H Sajjad, N Durrani, F Dalvi - Transactions of the Association for ..., 2022 - direct.mit.edu

The proliferation of **Deep** Neural Networks in various domains has seen an increased need for interpretability of these models. Preliminary work done along this line, and papers that ...

☆ Save Cite Cited by 8 Related articles All 2 versions

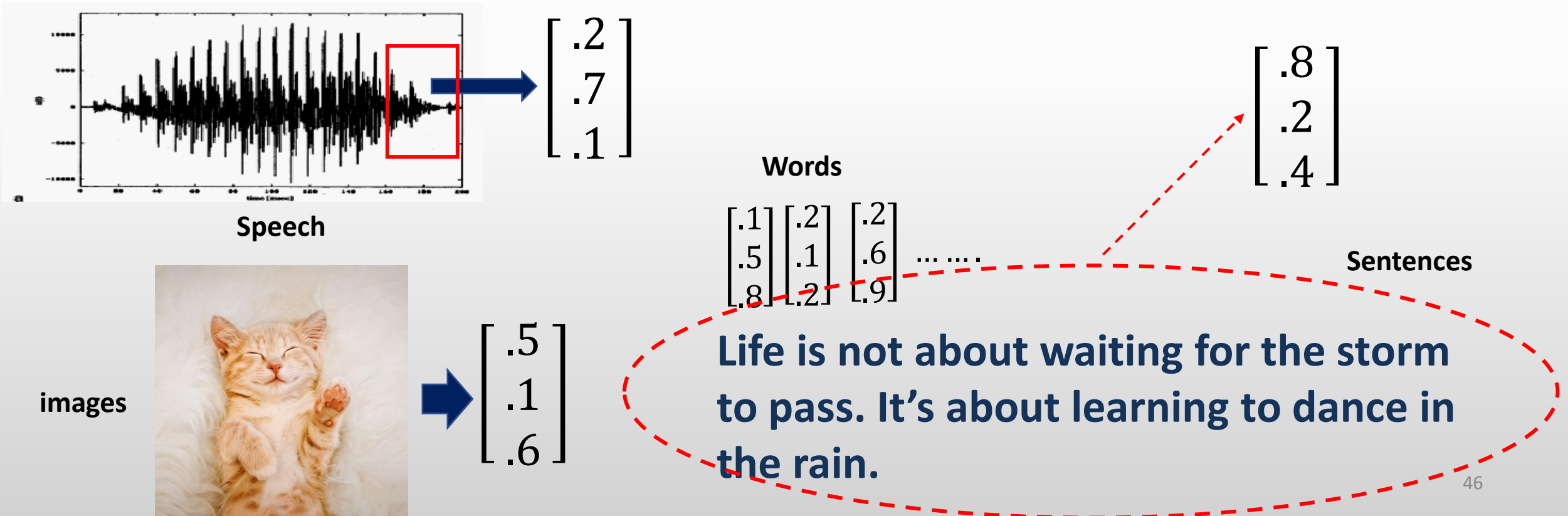


## Deep Learning-based NLP



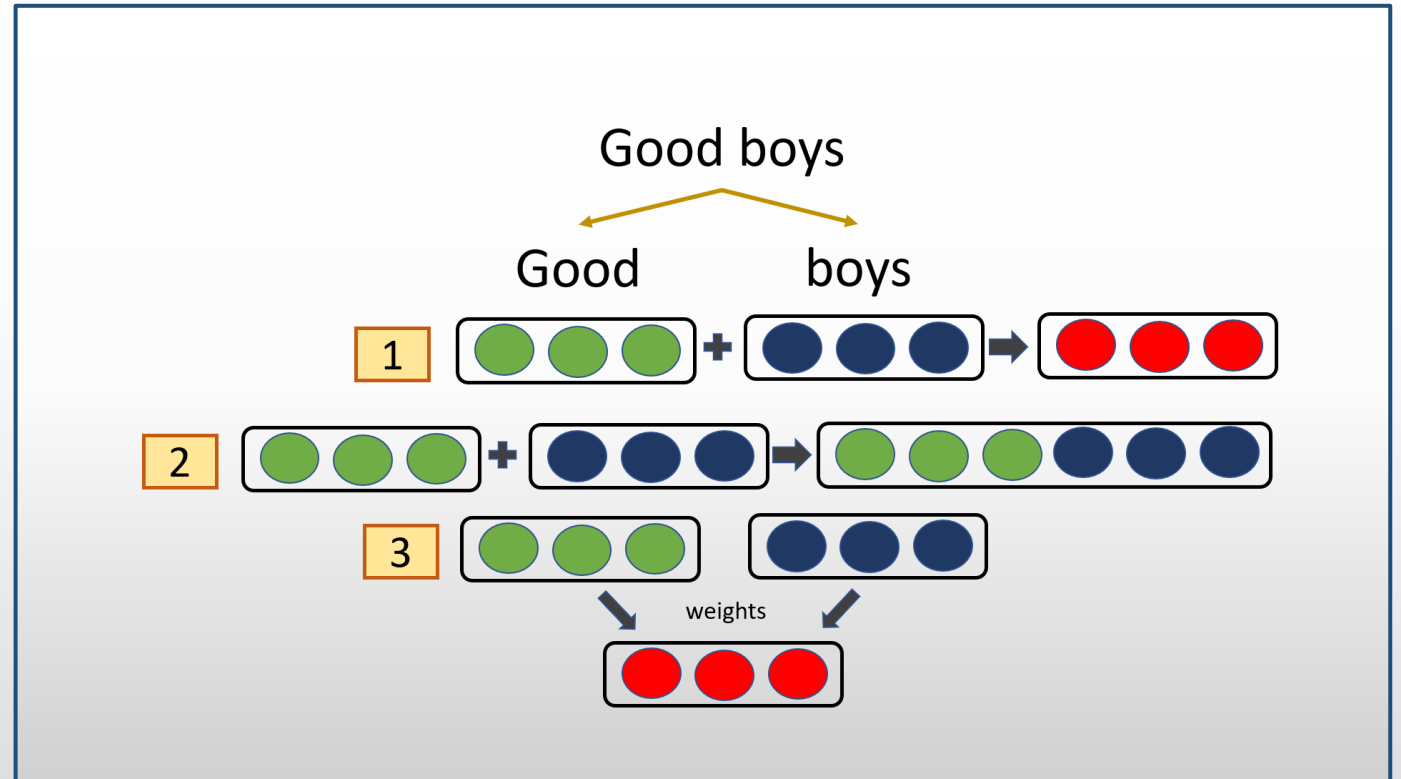
# Reasons for exploring Deep Learning

- Learned features are easy to adapt, fast to learn.
- Representing word, visual, and linguistic information as latent vector.



# Reasons for exploring Deep Learning (cont.)

- Flexible
- Can represent any levels of NLP
  - Word
  - Phrase
  - Sentence
  - Paragraph



# NLP tools



# Framework & Implementation



# NLP Libraries

NLTK

Documentation

Search

Natural Language Toolkit

NLTK Documentation  
API Reference  
Example Usage

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.



<https://www.nltk.org/>

**spaCy** Out now: spaCy v3.4

USAGE MODELS API UNIVERSE 25,008 Search docs

## Industrial-Strength Natural Language Processing

IN PYTHON

### Get things done

spaCy is designed to help you do real work — to build real products, or gather real insights. The library respects your time, and tries to avoid wasting it. It's easy to install, and its API is simple and productive.

GET STARTED

### Blazing fast

spaCy excels at large-scale information extraction tasks. It's written from the ground up in carefully memory-managed Cython. If your application needs to process entire web dumps, spaCy is the library you want to be using.

FACTS & FIGURES

### Awesome ecosystem

In the five years since its release, spaCy has become an industry standard with a huge ecosystem. Choose from a variety of plugins, integrate with your machine learning stack and build custom components and workflows.

READ MORE

<https://spacy.io/>



Hugging Face

Search models,

• Course ▾



Search documentation

Ctrl+K

MAIN ▾

EN ▾



Computer Vision

- Image Classification
- Image Segmentation
- Zero-Shot Image Classification
- Image-to-Image
- Unconditional Image Generation
- Object Detection
- Video Classification
- Depth Estimation

Natural Language Processing

- Translation
- Fill-Mask
- Token Classification
- Sentence Similarity
- Question Answering
- Summarization
- Zero-Shot Classification
- Text Classification
- Text2Text Generation
- Text Generation
- Conversational
- Table Question Answering

Audio

- Automatic Speech Recognition
- Audio Classification
- Text-to-Speech
- Audio-to-Audio
- Voice Activity Detection

Multimodal

- Feature Extraction
- Text-to-Image
- Visual Question Answering
- Image-to-Text
- Document Question Answering

0. SETUP

1. TRANSFORMER MODELS

Introduction

Natural Language Processing

Transformers, what can they do?

How do Transformers work?

Encoder models

Decoder models

Sequence-to-sequence models

Bias and limitations

Summary

End-of-chapter quiz

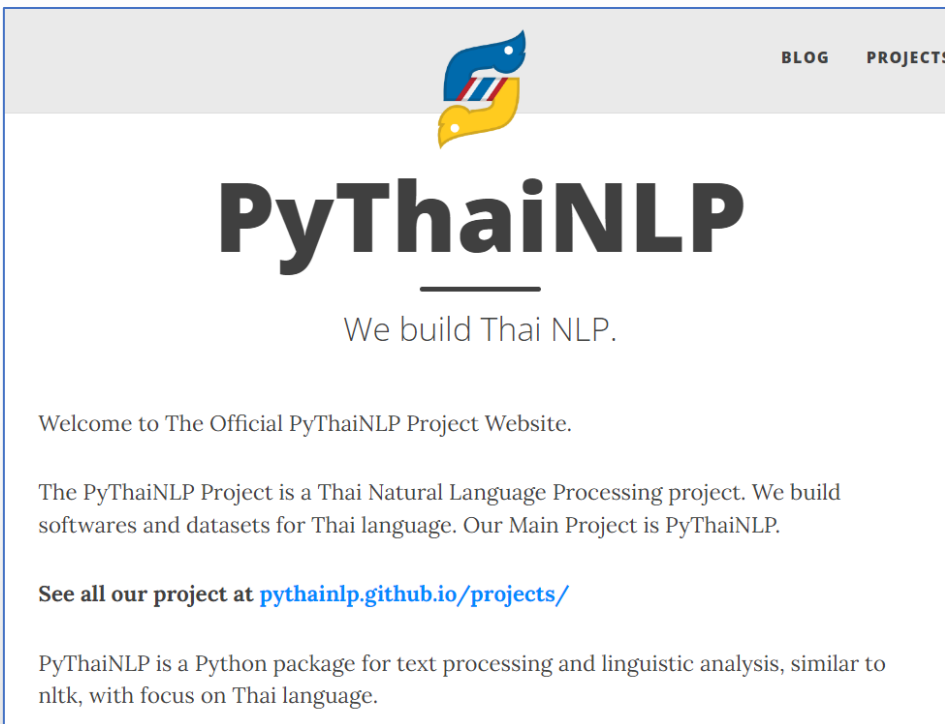
2. USING 🤖 TRANSFORMERS

3. FINE-TUNING A PRETRAINED MODEL

4. SHARING MODELS AND TOKENIZERS

<https://huggingface.co/>

# NLP Library for Thai



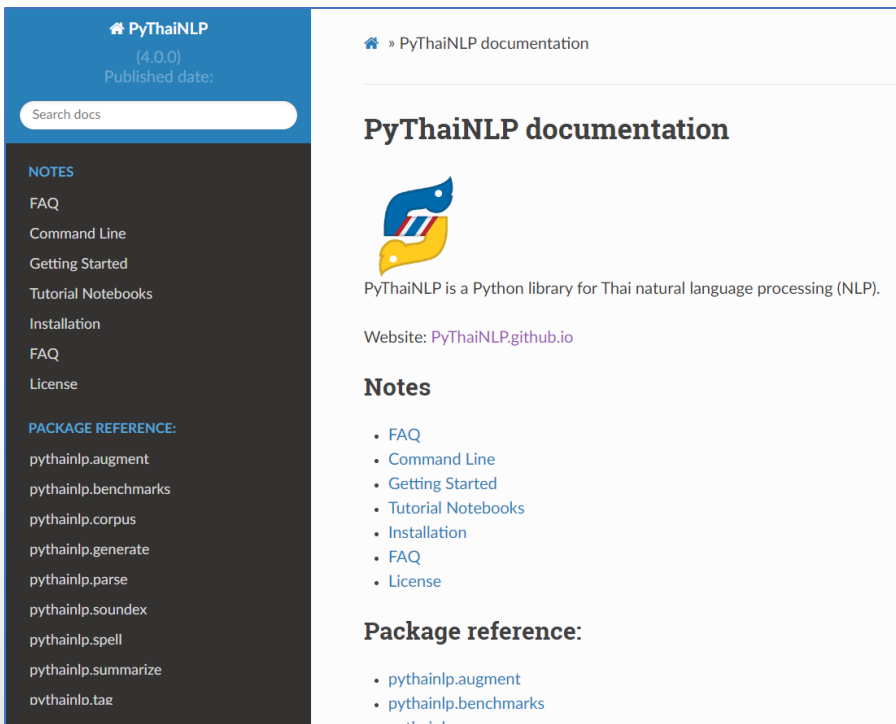
The image shows the official PyThaiNLP Project Website. It features the PyThaiNLP logo at the top, which consists of a stylized blue and yellow bird-like shape. Below the logo, the text "PyThaiNLP" is prominently displayed in a large, bold, black font. Underneath this, the tagline "We build Thai NLP." is written in a smaller, black font. The website has a clean, modern design with a light gray background. In the top right corner, there are links for "BLOG" and "PROJECTS". The main content area includes a welcome message, a description of the project, and a link to the GitHub repository.

Welcome to The Official PyThaiNLP Project Website.

The PyThaiNLP Project is a Thai Natural Language Processing project. We build softwares and datasets for Thai language. Our Main Project is PyThaiNLP.

See all our project at [pythainlp.github.io/projects/](https://pythainlp.github.io/projects/)

PyThaiNLP is a Python package for text processing and linguistic analysis, similar to nltk, with focus on Thai language.



The image shows the PyThaiNLP documentation page. It has a blue header with the PyThaiNLP logo and the version number (4.0.0). Below the header, there is a search bar and a sidebar with navigation links. The main content area is titled "PyThaiNLP documentation" and includes the PyThaiNLP logo, a description of the library, the website link, and a list of notes. The sidebar contains links to various sections like NOTES, FAQ, Command Line, Getting Started, Tutorial Notebooks, Installation, FAQ, and License. The main content area also includes a "Package reference" section with links to various PyThaiNLP packages.

PyThaiNLP (4.0.0)  
Published date:

Search docs

**PyThaiNLP documentation**

PyThaiNLP is a Python library for Thai natural language processing (NLP).

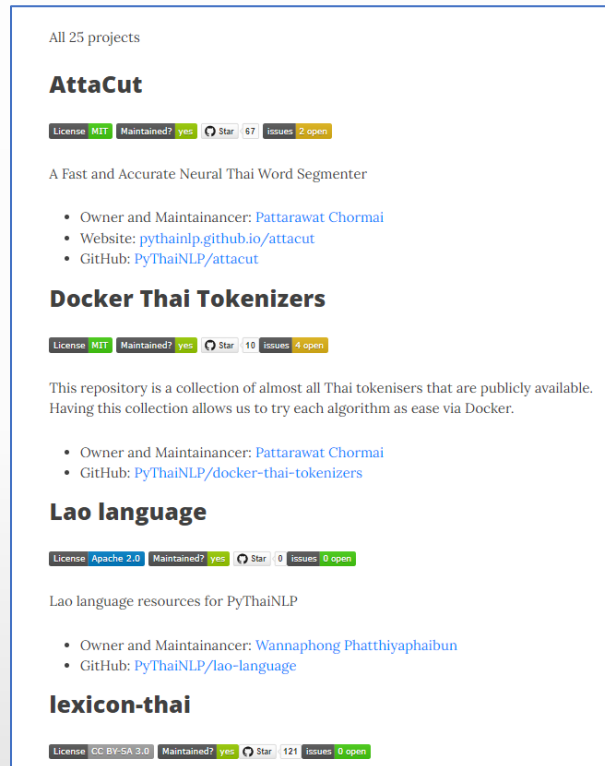
Website: [PyThaiNLP.github.io](https://pythainlp.github.io)

**Notes**

- [FAQ](#)
- [Command Line](#)
- [Getting Started](#)
- [Tutorial Notebooks](#)
- [Installation](#)
- [FAQ](#)
- [License](#)

**Package reference:**

- [pythainlp.augment](#)
- [pythainlp.benchmarks](#)
- [pythainlp.corpus](#)



The image shows a GitHub repository page for PyThaiNLP projects. It lists three repositories: AttaCut, Docker Thai Tokenizers, and Lao language. Each repository has a license, a maintainer, a website, and a GitHub link. The AttaCut repository is a Fast and Accurate Neural Thai Word Segmenter. The Docker Thai Tokenizers repository is a collection of almost all Thai tokenisers that are publicly available. The Lao language repository is a collection of Lao language resources for PyThaiNLP.

All 25 projects

**AttaCut**

License: MIT Maintained? yes Star: 67 Issues: 2 open

A Fast and Accurate Neural Thai Word Segmenter

- Owner and Maintainer: [Pattarawat Chormai](#)
- Website: [pythainlp.github.io/attacut](https://pythainlp.github.io/attacut)
- GitHub: [PyThaiNLP/attacut](https://github.com/PyThaiNLP/attacut)

**Docker Thai Tokenizers**

License: MIT Maintained? yes Star: 10 Issues: 4 open

This repository is a collection of almost all Thai tokenisers that are publicly available. Having this collection allows us to try each algorithm as ease via Docker.

- Owner and Maintainer: [Pattarawat Chormai](#)
- GitHub: [PyThaiNLP/docker-thai-tokenizers](https://github.com/PyThaiNLP/docker-thai-tokenizers)

**Lao language**

License: Apache 2.0 Maintained? yes Star: 0 Issues: 0 open

Lao language resources for PyThaiNLP

- Owner and Maintainer: [Wannaphong Phatthiyaphaibun](#)
- GitHub: [PyThaiNLP/lao-language](https://github.com/PyThaiNLP/lao-language)

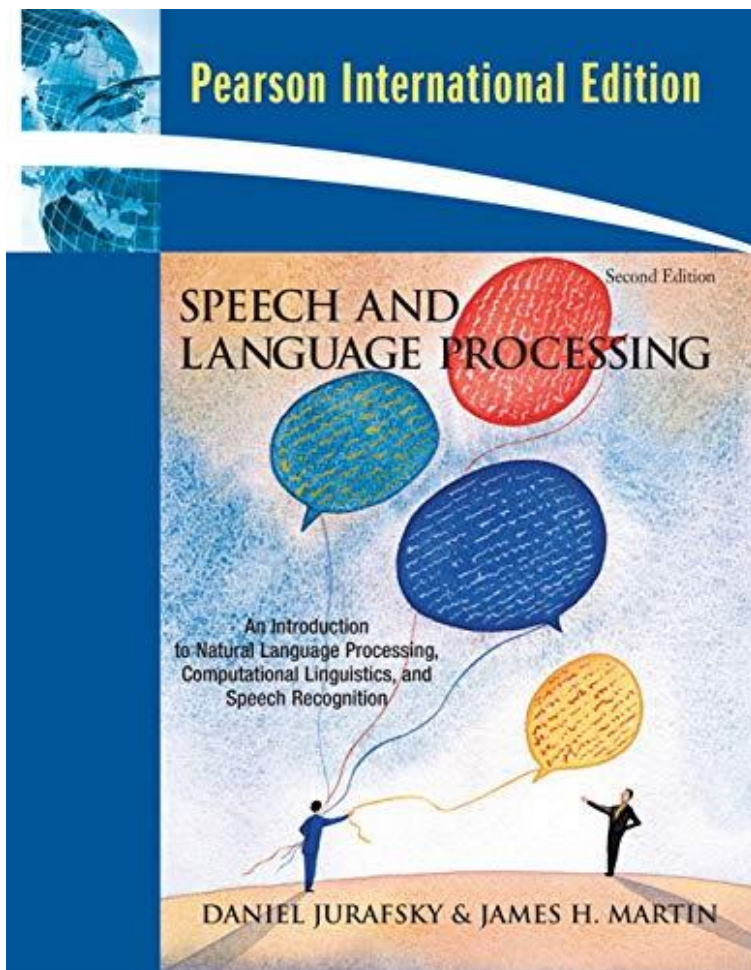
**lexicon-thai**

License: CC-BY-SA 3.0 Maintained? yes Star: 121 Issues: 0 open

<https://pythainlp.github.io/>  
<https://pythainlp.github.io/docs/2.0/index.html>

# External sources

- Speech and Language Processing (3rd ed. draft)
  - <https://web.stanford.edu/~jurafsky/slp3/>



Chapter	Slides
<b>Part I: Fundamental Algorithms</b>	
1: Introduction	
2: <a href="#">Regular Expressions, Text Normalization, Edit Distance</a>	2: Text Processing [ <a href="#">pptx</a> ] [ <a href="#">pdf</a> ] 2: Edit Distance [ <a href="#">pptx</a> ] [ <a href="#">pdf</a> ]
3: <a href="#">N-gram Language Models</a>	3: N-grams [ <a href="#">pptx</a> ] [ <a href="#">pdf</a> ]
4: <a href="#">Naive Bayes and Sentiment Classification</a>	4: Naive Bayes + Sentiment [ <a href="#">pptx</a> ] [ <a href="#">pdf</a> ]
5: <a href="#">Logistic Regression</a>	5: LR [ <a href="#">pptx</a> ] [ <a href="#">pdf</a> ]
6: <a href="#">Vector Semantics and Embeddings</a>	6: Vector Semantics [ <a href="#">pptx</a> ] [ <a href="#">pdf</a> ]
7: <a href="#">Neural Networks and Neural Language Models</a>	7: Neural Networks [ <a href="#">pptx</a> ] [ <a href="#">pdf</a> ]
8: <a href="#">Sequence Labeling for Parts of Speech and Named Entities</a>	8: POS/NER Intro only [ <a href="#">pptx</a> ] [ <a href="#">pdf</a> ]
9: <a href="#">RNNs and LSTMs</a>	
10: <a href="#">Transformers and Pretrained Language Models</a>	
11: <a href="#">Fine-tuning and Masked Language Models</a>	
12: Prompting and Instruct Tuning	
<b>Part II: NLP Applications</b>	
13: <a href="#">Machine Translation</a>	
14: <a href="#">Question Answering and Information Retrieval</a>	
15: <a href="#">Chatbots and Dialogue Systems</a>	15: Dialog [ <a href="#">pptx</a> ] [ <a href="#">pdf</a> ]
16: <a href="#">Automatic Speech Recognition and Text-to-Speech</a>	