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

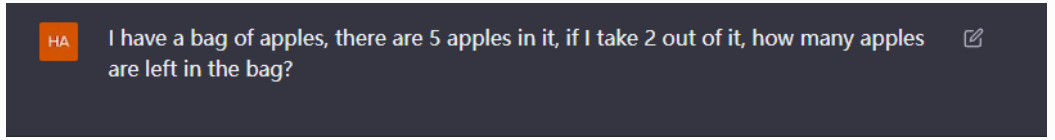
Outlines

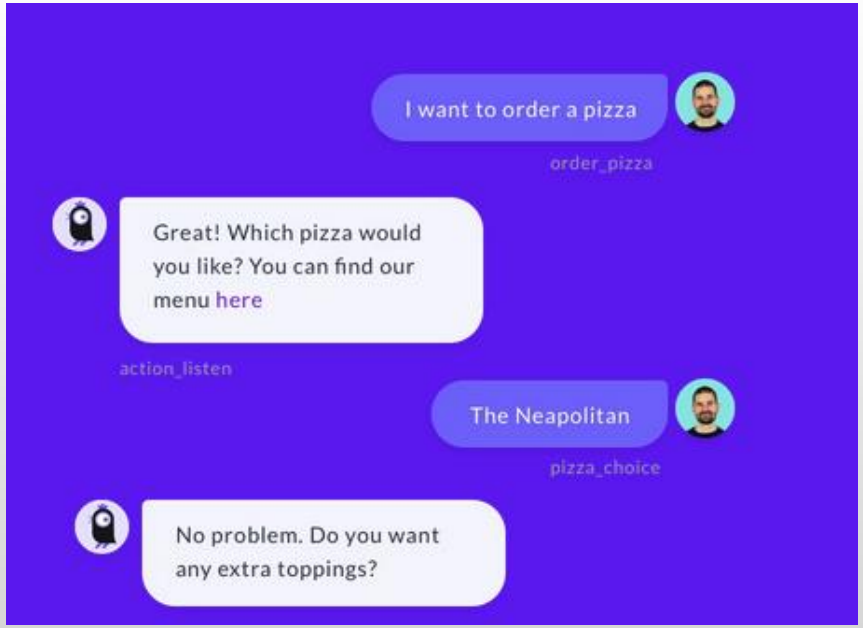
- What is Natural Language Processing (NLP) ?
- Why it is difficult ?
- Processing steps of NLP
- Deep Learning in NLP
- NLP tools

What is Natural Language Processing (NLP) ?

Natural Language Processing (NLP)

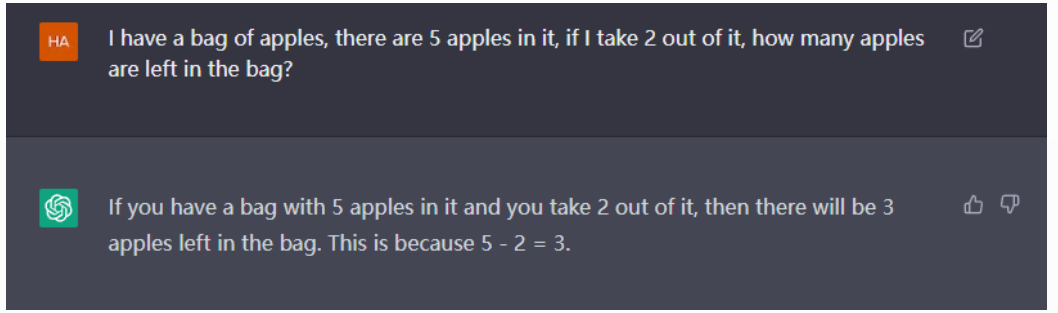
- NLP is a subfield in AI, where the goal is
  - To bright the gap between how people communicate and what machines understand in order to perform useful tasks, e.g. making appointments, translation, question answering (QA), etc.

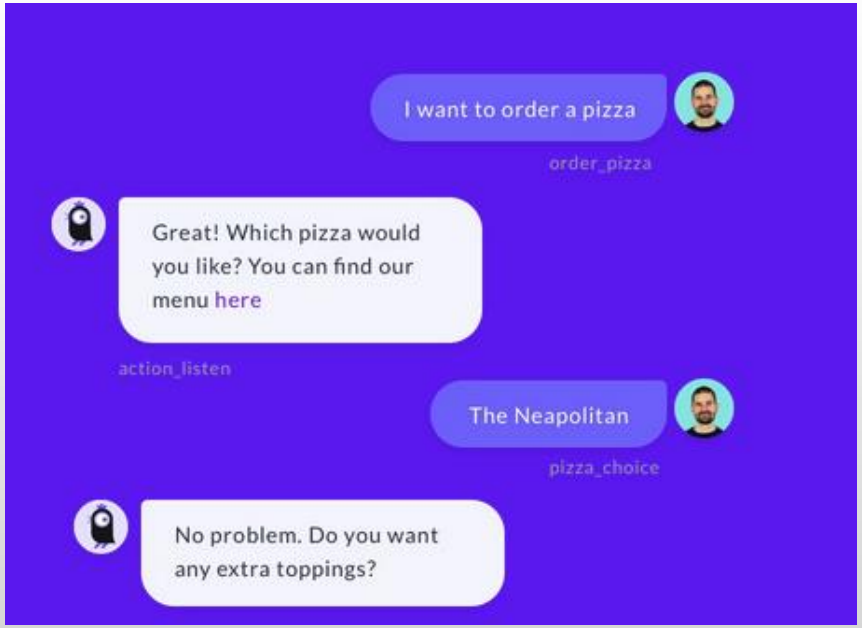




Natural Language Processing (NLP)

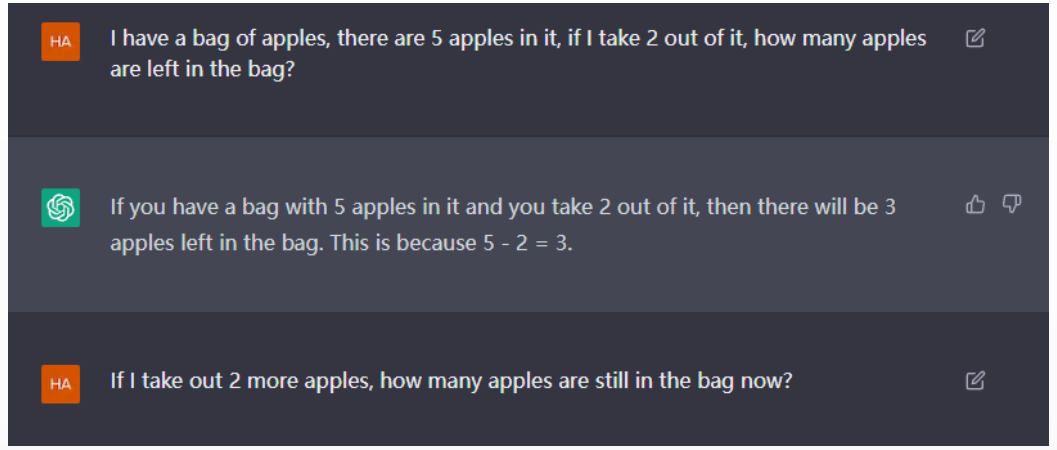
- NLP is a subfield in AI, where the goal is
  - To bright the gap between how people communicate and what machines understand in order to perform useful tasks, e.g. making appointments, translation, question answering (QA), etc.

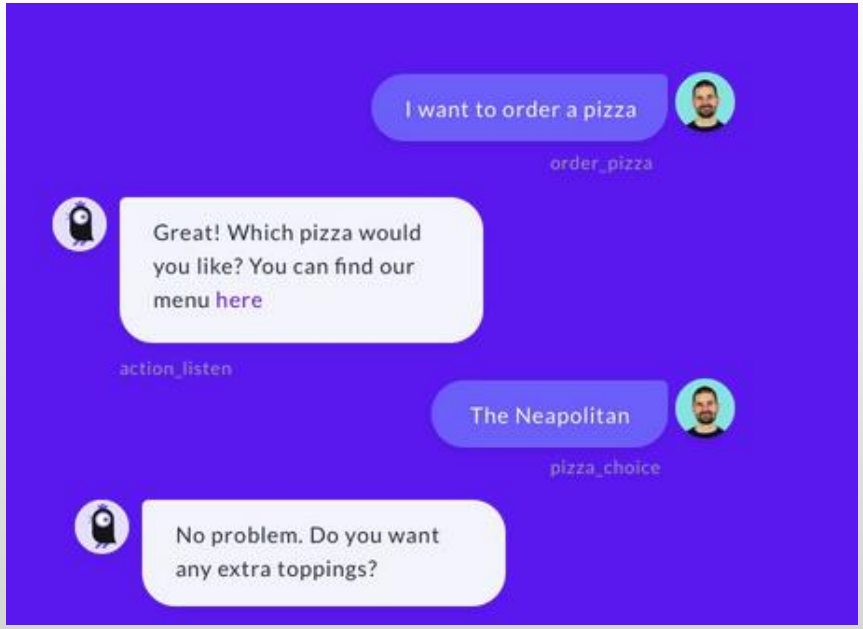




Natural Language Processing (NLP)

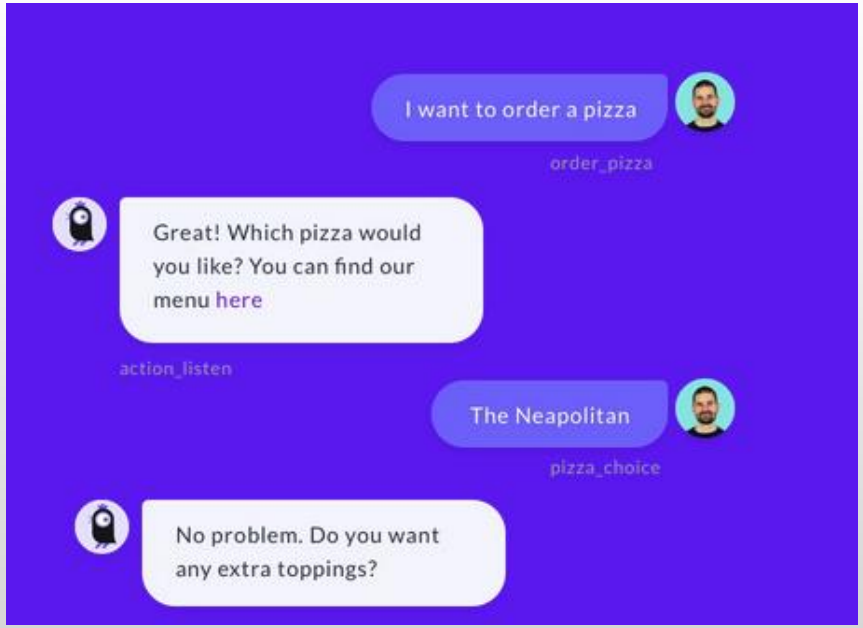
- NLP is a subfield in AI, where the goal is
  - To bright the gap between how people communicate and what machines understand in order to perform useful tasks, e.g. making appointments, translation, question answering (QA), etc.

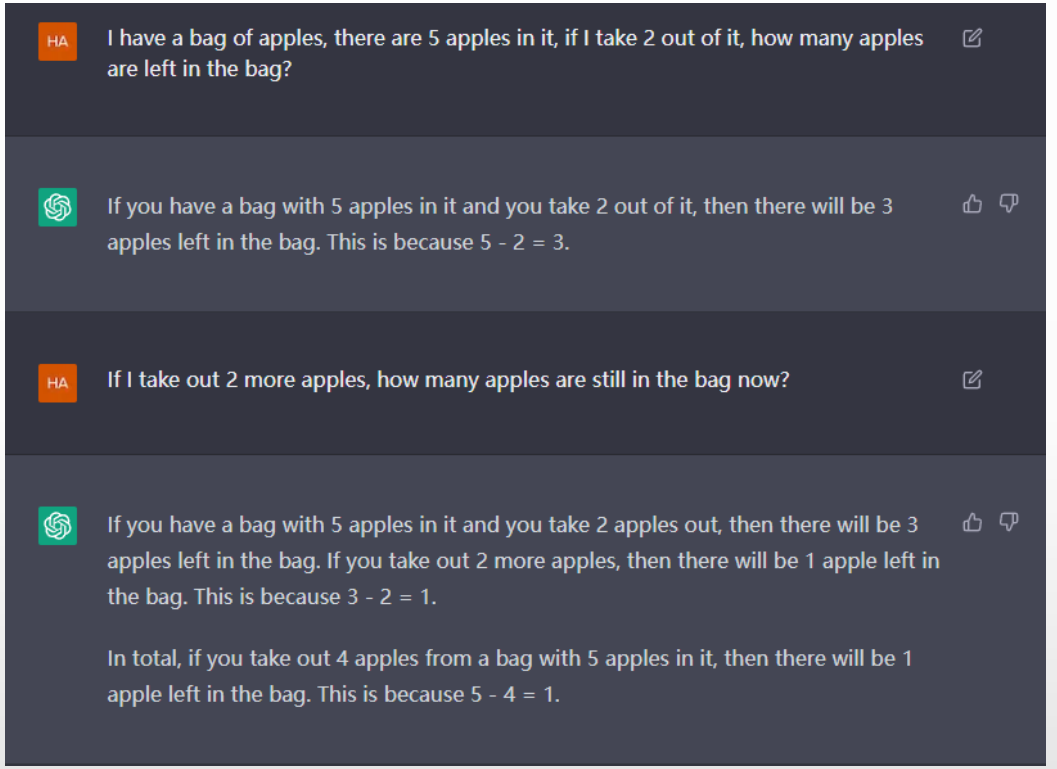




Natural Language Processing (NLP)

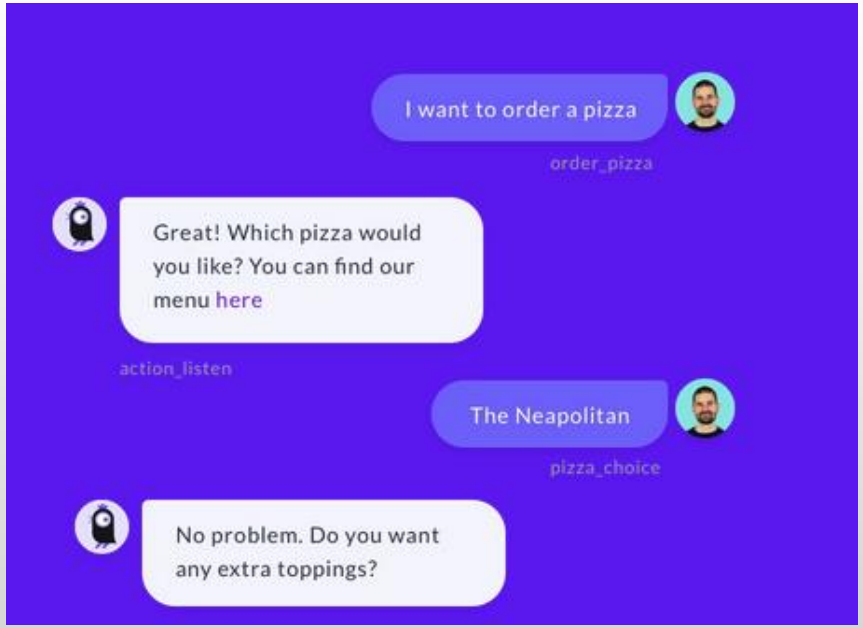
- NLP is a subfield in AI, where the goal is
  - To bright the gap between how people communicate and what machines understand in order to perform useful tasks, e.g. making appointments, translation, question answering (QA), etc.

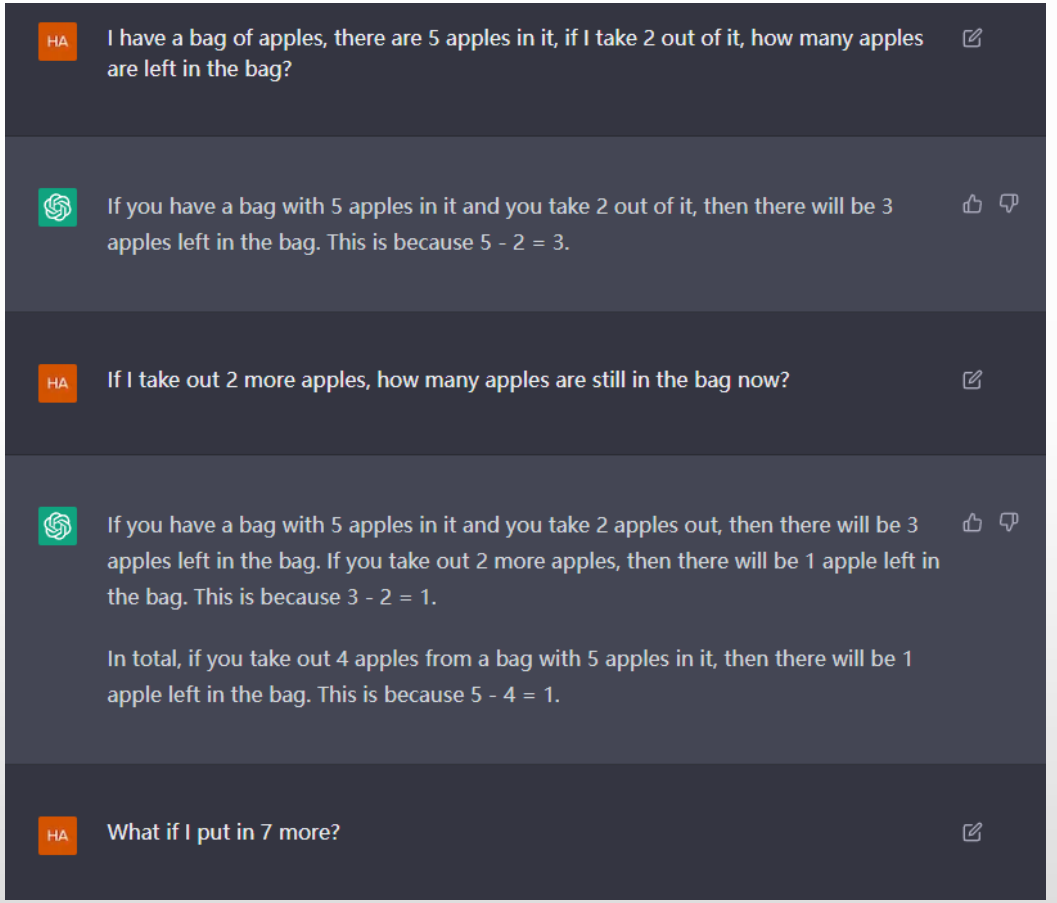




Natural Language Processing (NLP)

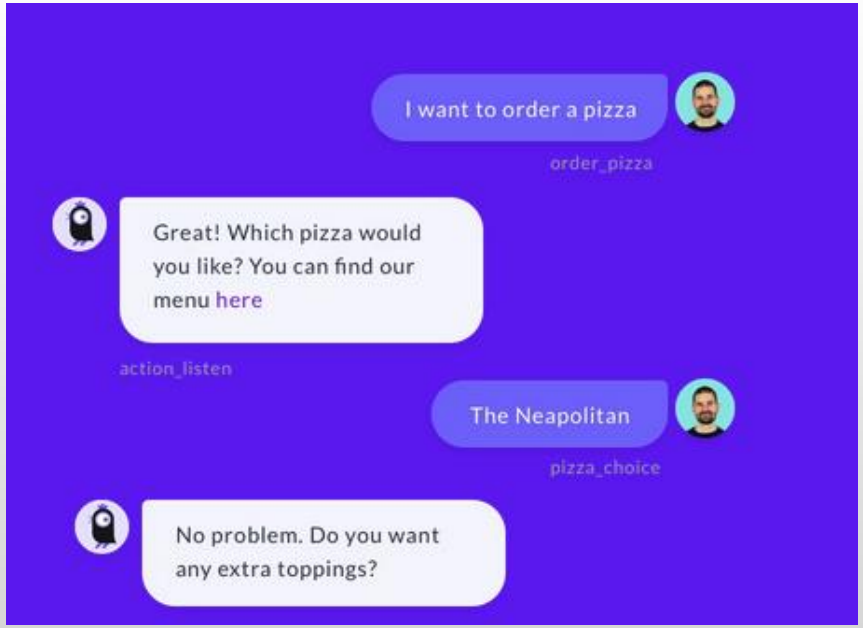
- NLP is a subfield in AI, where the goal is
  - To bright the gap between how people communicate and what machines understand in order to perform useful tasks, e.g. making appointments, translation, question answering (QA), etc.





Natural Language Processing (NLP)

- NLP is a subfield in AI, where the goal is
  - To bright 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.



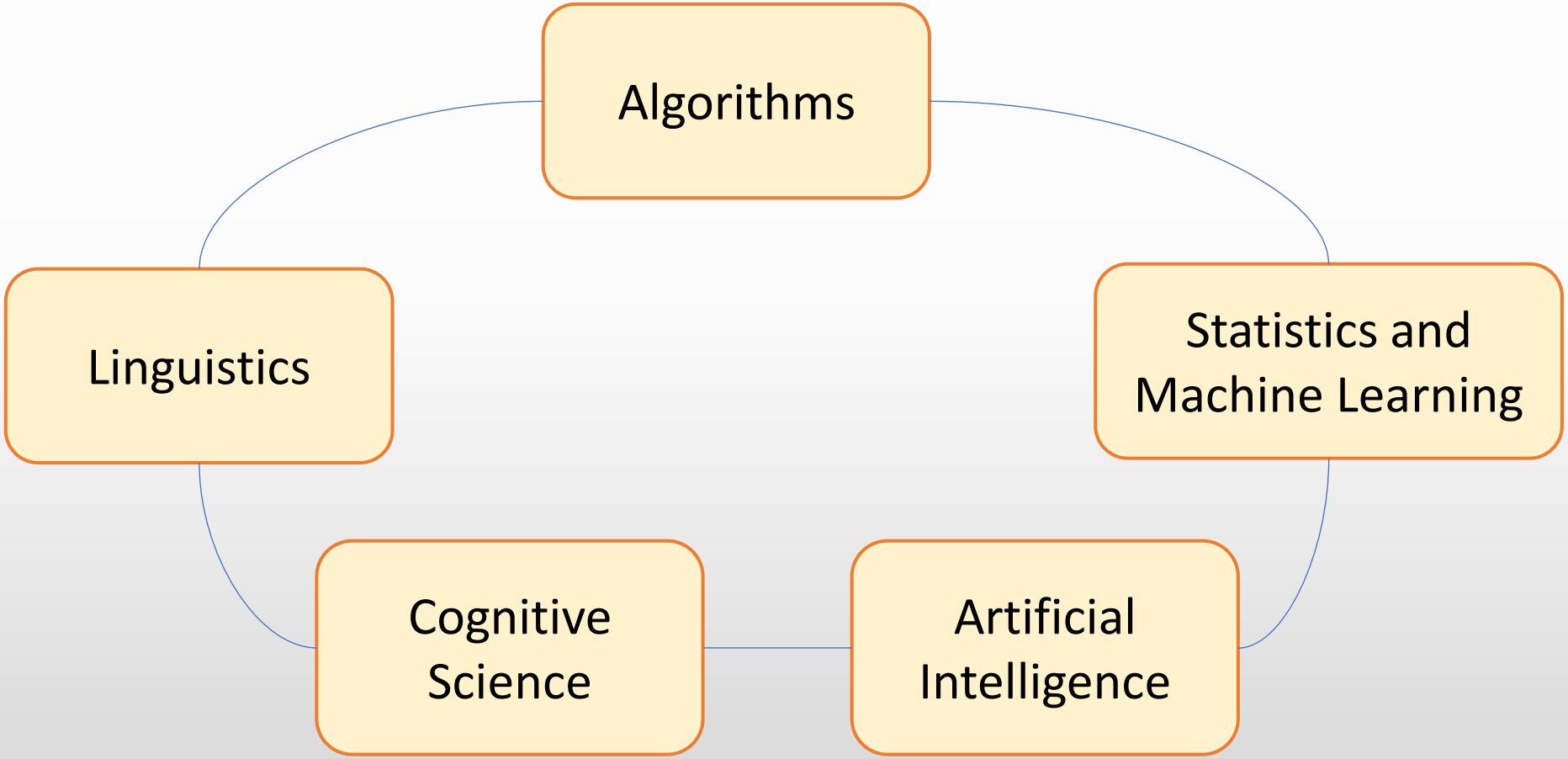
NLP Goals

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

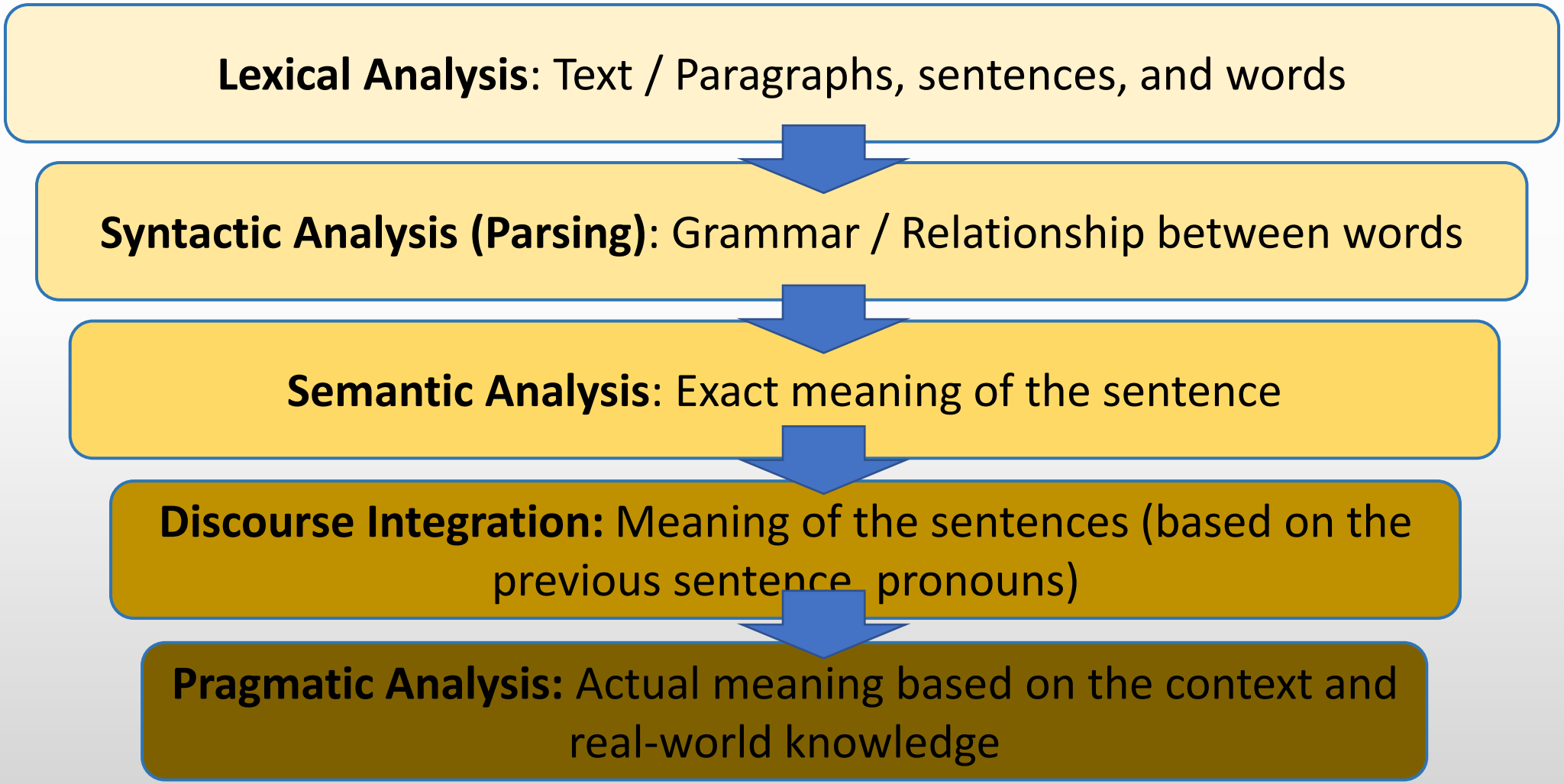
NLP Goals

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

NLP is interdisciplinary

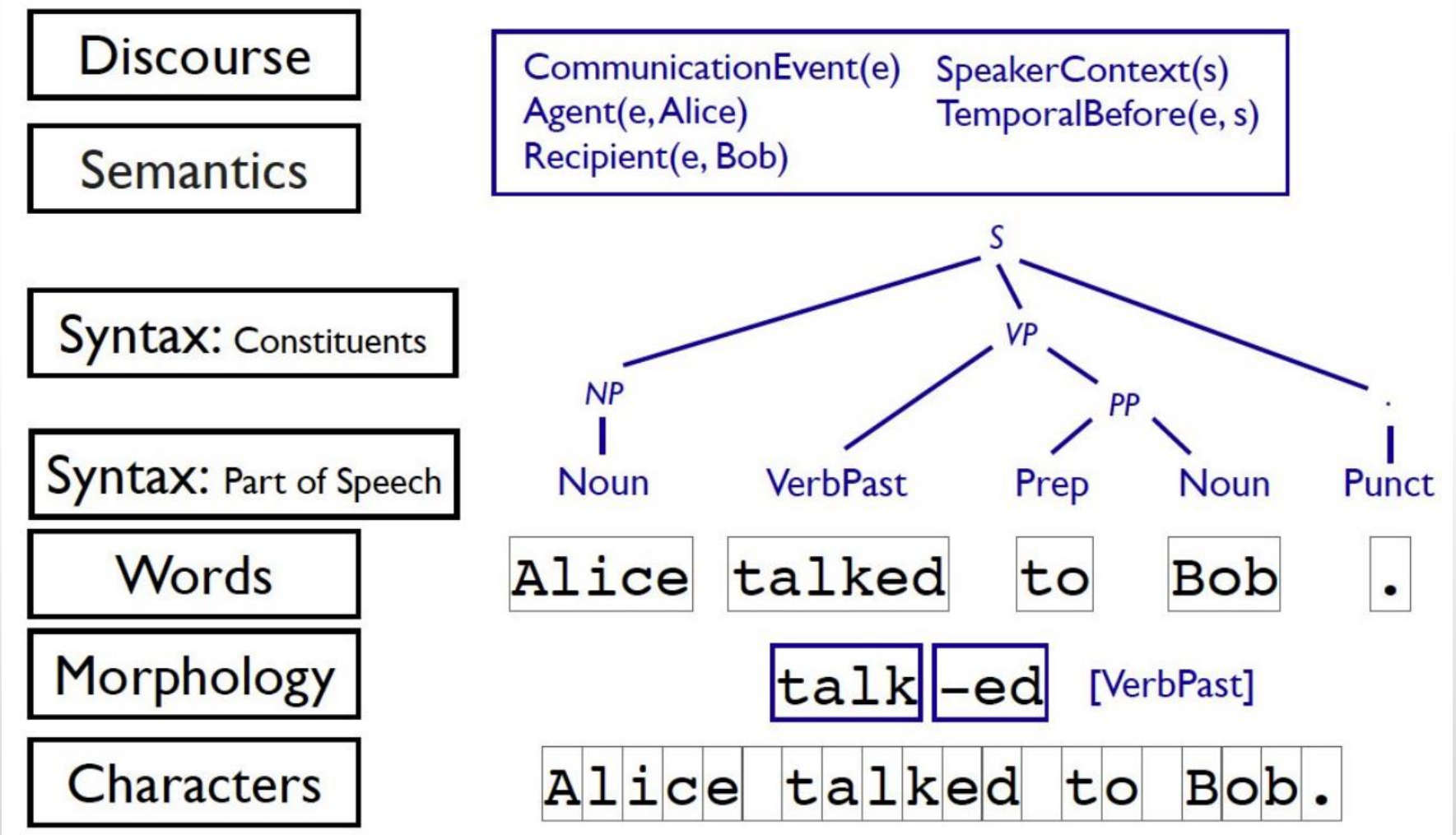


Level of understand in NLP



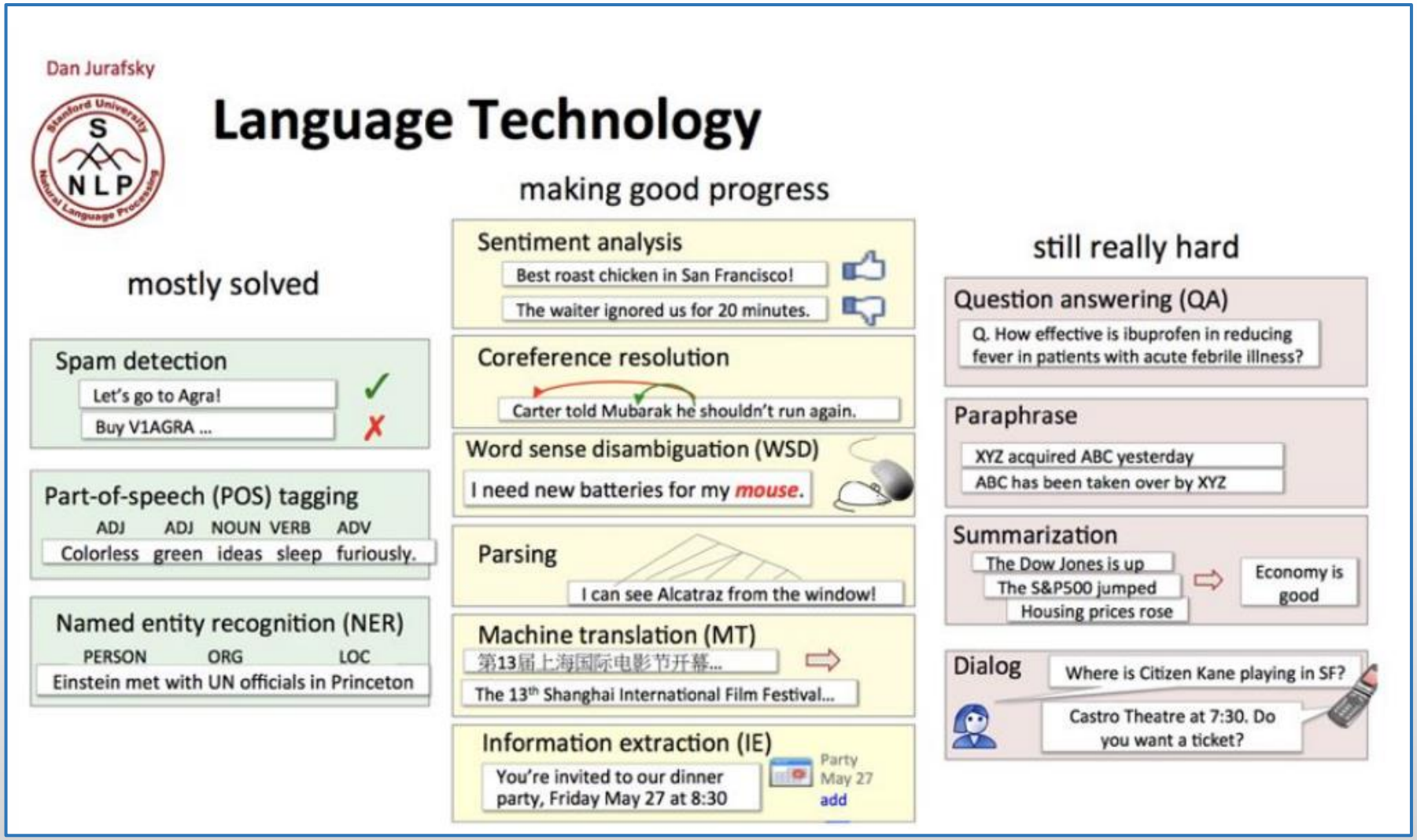
https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_natural_language_processing.htm

Level of understand in NLP (cont.)

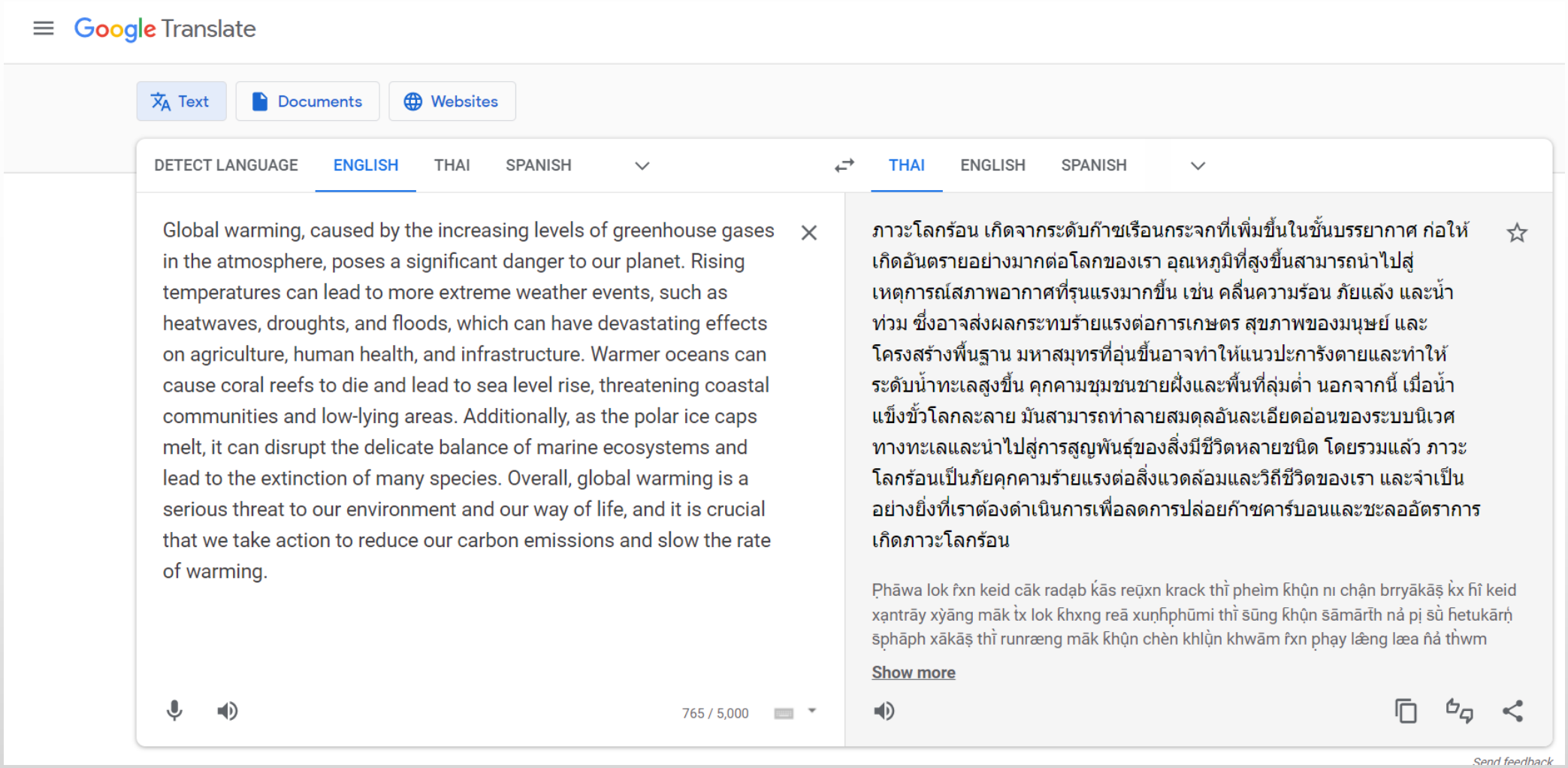


https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_natural_language_processing.htm

NLP today and Technology



NLP today: Machine Translation (MT)



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/

ChatGPT

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?

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?

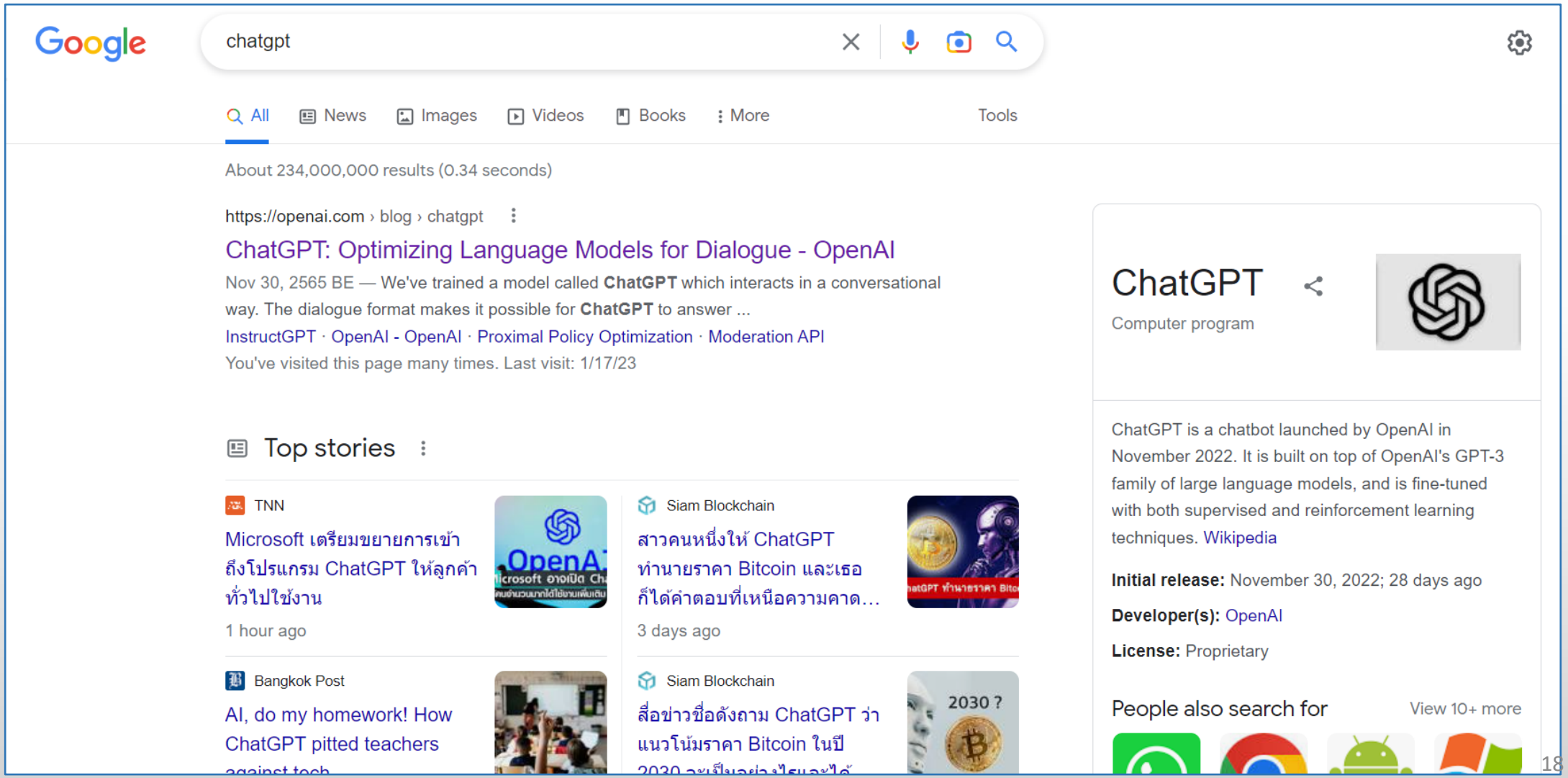
If you have a bag with 5 apples in it and you take 2 apples out, then there will be 3 apples left in the bag. If you take out 2 more apples, then there will be 1 apple left in the bag. This is because 3 - 2 = 1.

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.

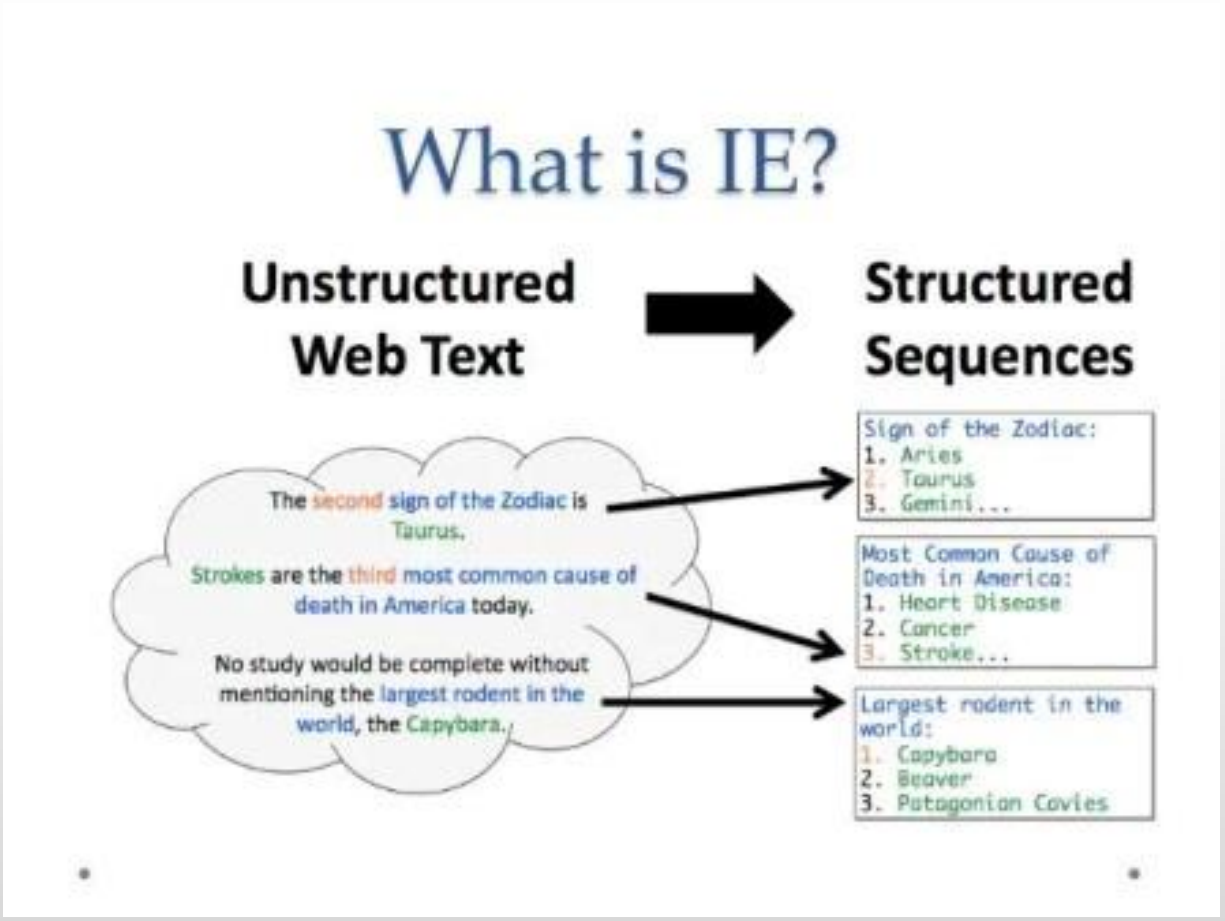
NLP today: Search & Summarization



NLP today: Information Extraction

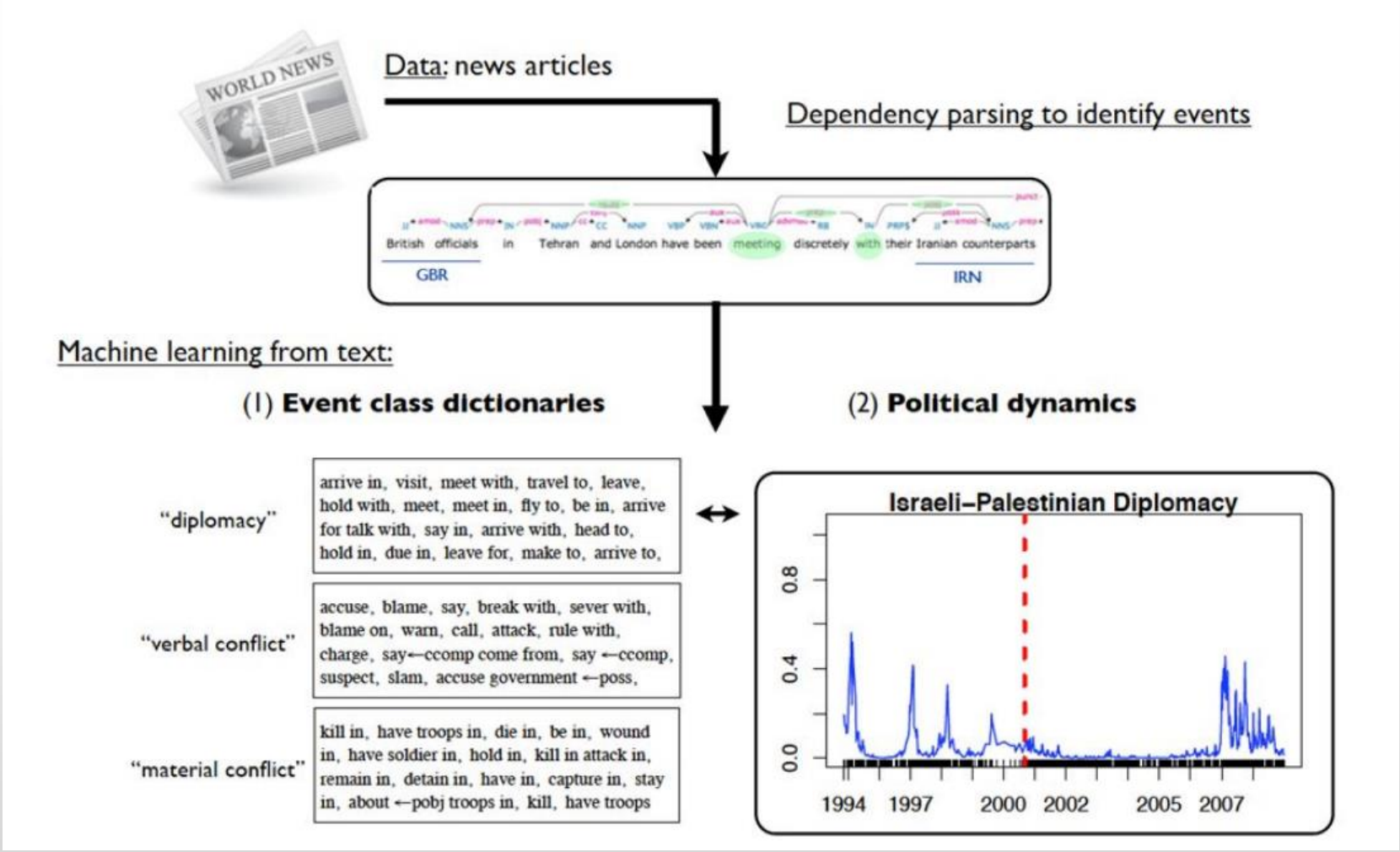
Abstract clinical report into a database

ID	AGE	RACE	STUDY	PROC	BIRTHS	MA_AGE	ASSESS	DENSITY	FINDING	FINDING T
9527	78	2	6/12/06	BIDXU-L	0		P	3	CALCS	N
32875	56	1	7/11/06	BIDXB-B	0		N	3		
2247	72	1	4/12/06	BIDXU-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	BIDXU-L	0		N	3		
67960	64	1	4/7/06	BIDXU-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		



https://www.engati.com/glossary/information-extraction

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 tl Internet about 'Hathaway' as the IMDb's StarMeter, and they're applying it to t stock market."

Ref: Prof. Regina Barzilay, NLP @MIT

NLP today: Trend analysis (cor

BERKSHIRE HATHAWAY

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 tl Internet about 'Hathaway' as the IMDb's StarMeter, and they're applying it to t 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 (cor

BERKSHIRE HATHAWAY

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 tl Internet about 'Hathaway' as the IMDb's StarMeter, and they're applying it to t stock market."

Ref: Prof. Regina Barzilay, NLP @MIT

NLP is difficult.
Word –level ambiguity !!!

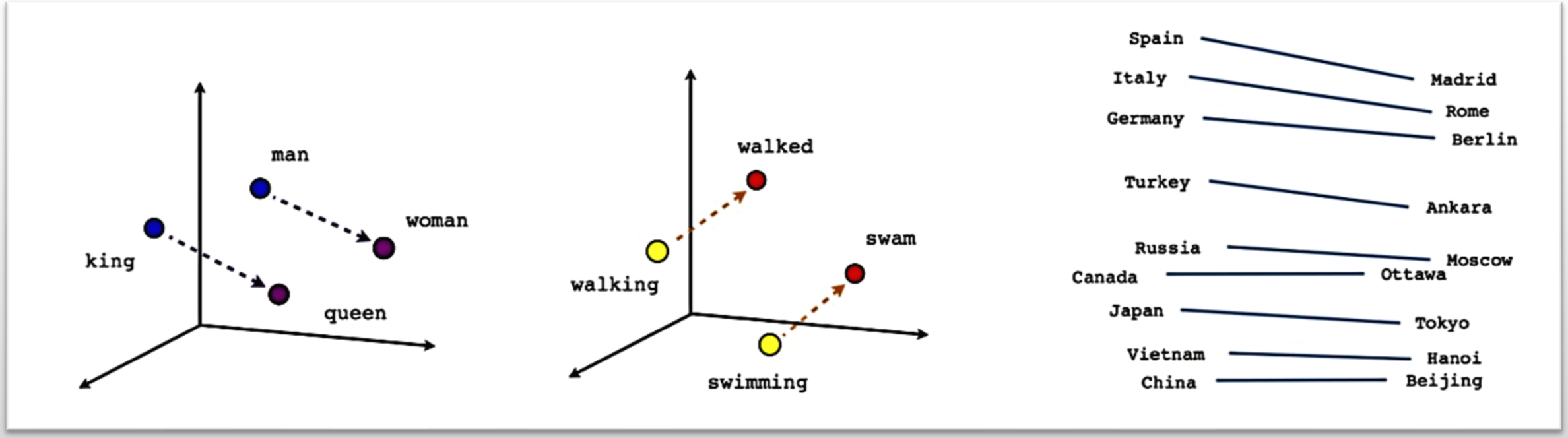
BERKSHIRE HATHAWAY INC.
3555 Farnam Street
Omaha, NE 68131
Official Home Page

- 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.)
- Corporate Governance
- Sustainability
- Common Stock Information
- Information Regarding Warren E. Buffett's Annual Contributions to Five Foundations
- Facts Regarding Berkshire's 2021 Investments in Activision Common Stock
- Berkshire Activewear

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.)

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

What NLP is difficult? (cont.)

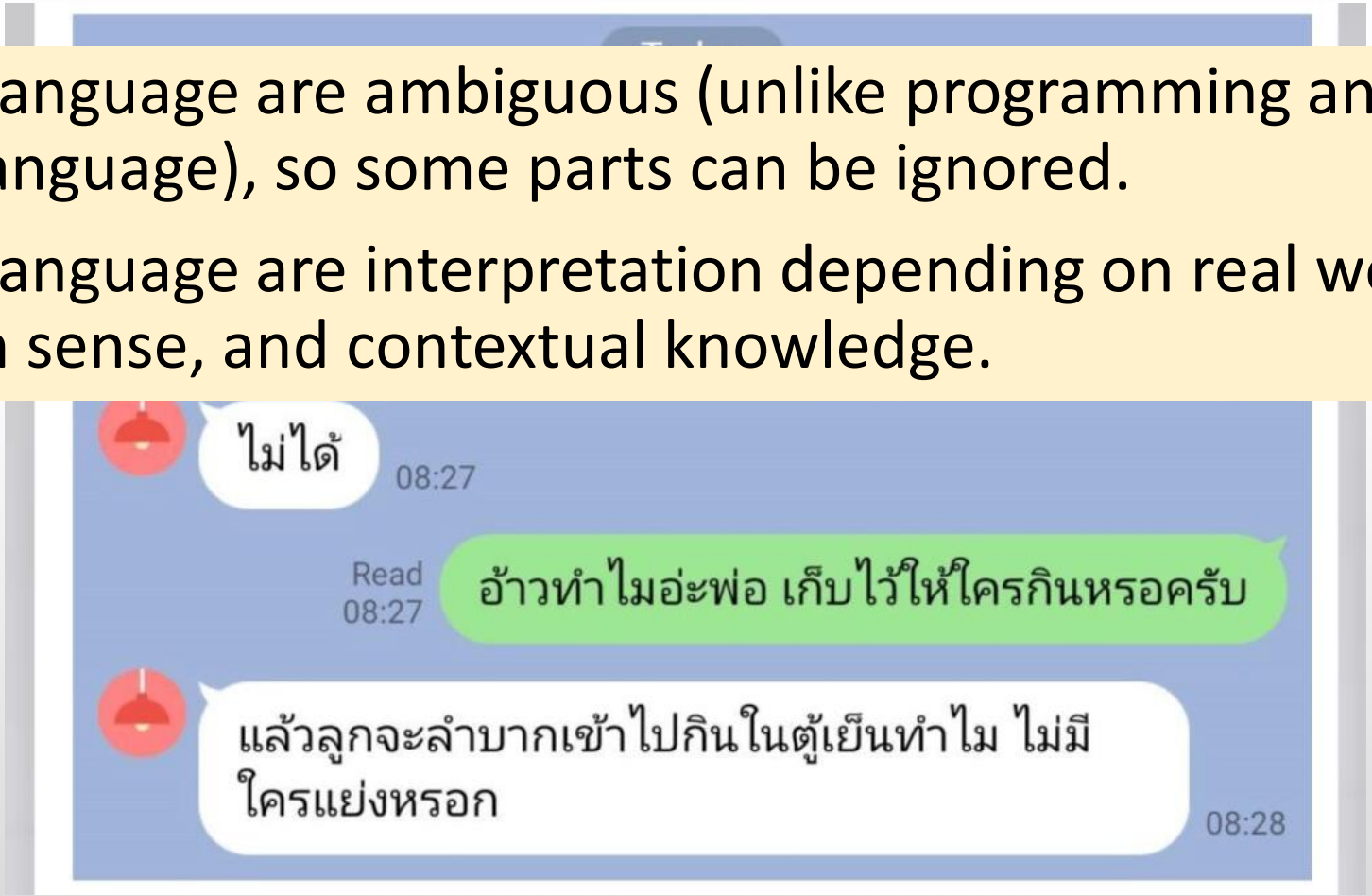
https://www.sti.chula.ac.th/knowledge/2820/

What NLP is difficult? (cont.)

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.



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

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|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|กำลัง|จะ|ไป|เรียน|วิชา|ประมวลผลภาษ
- ฉัน|กำลังจะ|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) เป็นแชทบอทปัญญาประดิษฐ์ต้นแบบที่พัฒนาโดยโอเพนเอไอ ซึ่งเชี่ยวชาญด้านการสนทนา แชทบอทเป็นโมเดลภาษาขนาดใหญ่ที่ได้รับการปรับแต่งอย่างละเอียดทั้งการควบคุมและเทคนิคการเรียนรู้แบบเสริมแรง 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 จะส่งผลกระทบต่อความต้องการของผู้ปฏิบัติงานด้านความรู้[1] ใน *เดอะเวิร์จ* เจมส์ วินเซนต์ มองเห็นความสำเร็จอย่างแพร่หลายของแชทจีพีทีเป็นหลักฐานว่าปัญญาประดิษฐ์กลายเป็นกระแสหลักไปแล้ว[2] ใน *ดิแอตแลนติก* สตีเฟน มาร์เช่ ตั้งข้อสังเกตว่ามีผลกระทบต่อสถาบันการศึกษาและโดยเฉพาะอย่างยิ่งเรียงความการสมัครงานยังไม่เป็นที่เข้าใจ[3] แดเนียล เฮอร์แมน ครูและนักเขียนของโรงเรียนมัธยมในแคลิฟอร์เนียเขียนว่า แชทจีพีทีจะนำเข้าสู่ "จุดจบของภาษาอังกฤษระดับมัธยมปลาย"[4]

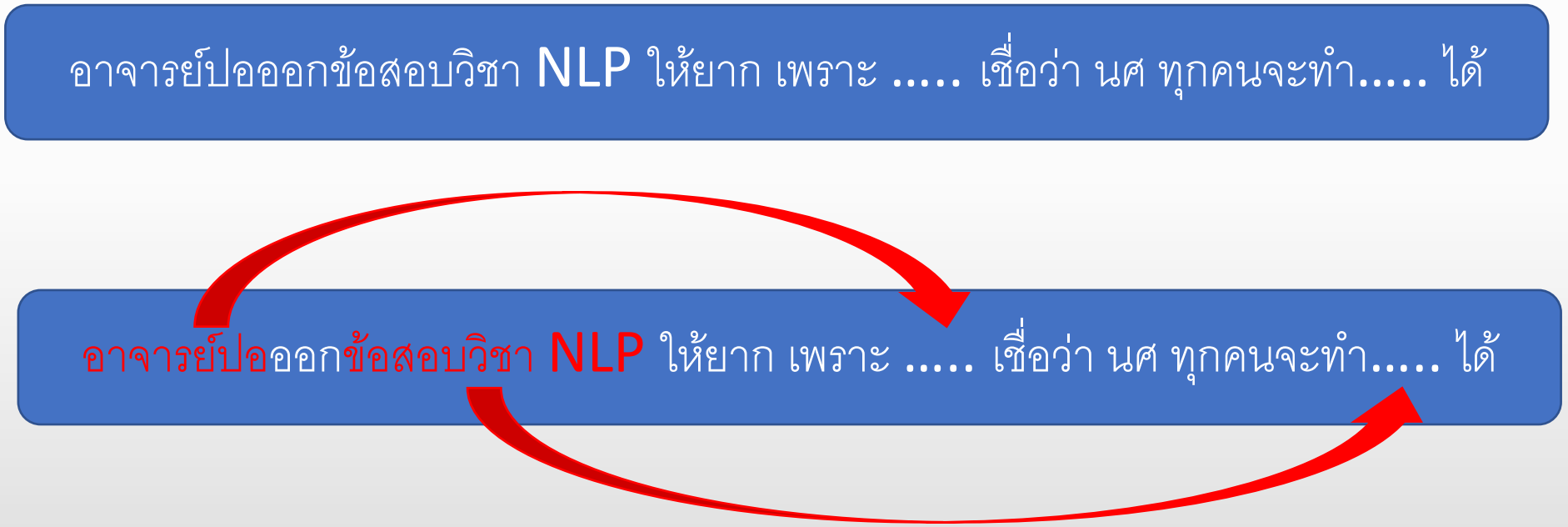
What NLP is difficult? (cont.)

- Syntax ambiguity

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

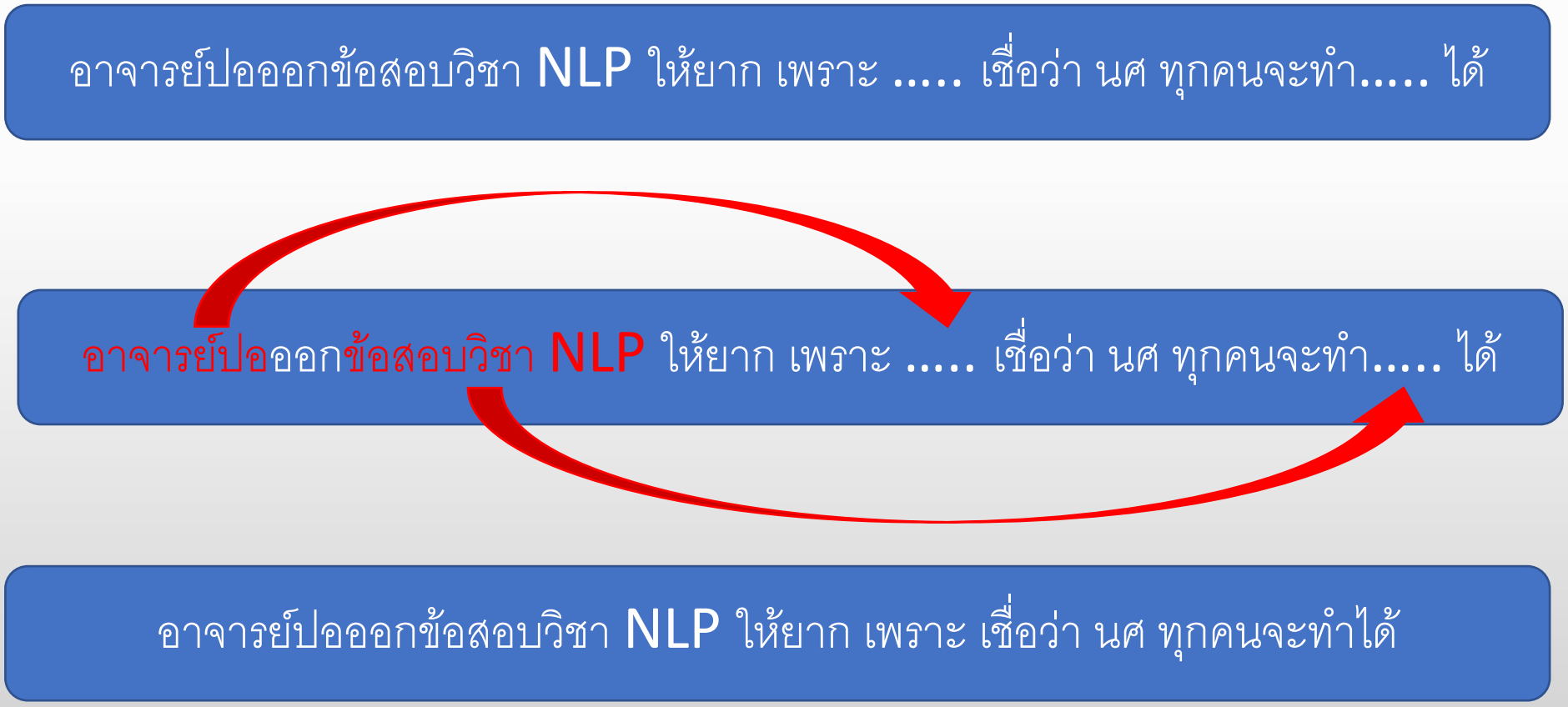
What NLP is difficult? (cont.)

- Syntax ambiguity



What NLP is difficult? (cont.)

- Syntax ambiguity



What NLP is difficult? (cont.)

- Slang

นร: ทำไมอาจารย์ออกข้อสอบวิชา NLP ยากจังครับ
อาจารย์ปอ: เพราะอาจารย์เลือดกรุ๊ปบี

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.

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)]

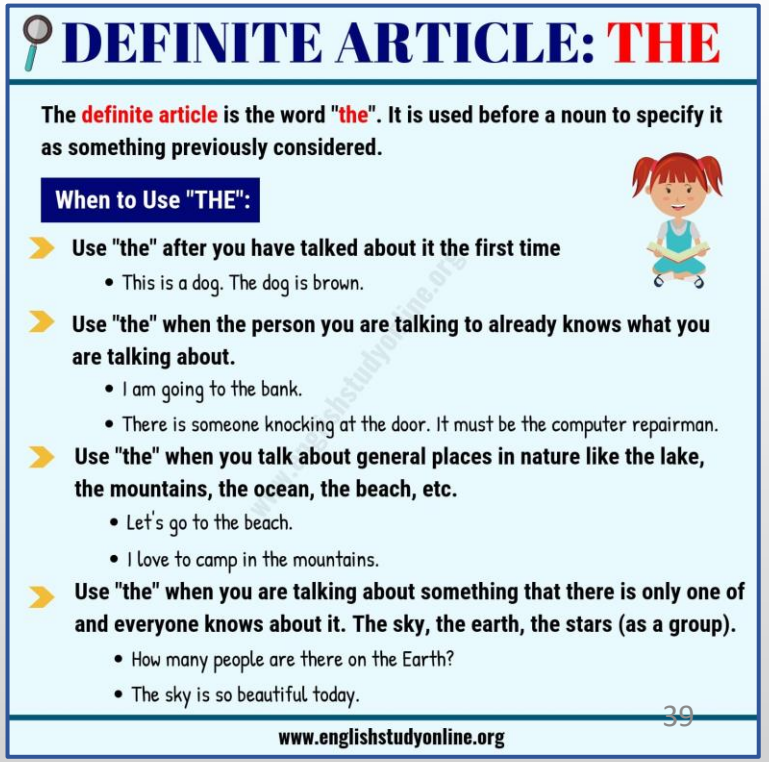
Application - e.g. Word Cloud

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.

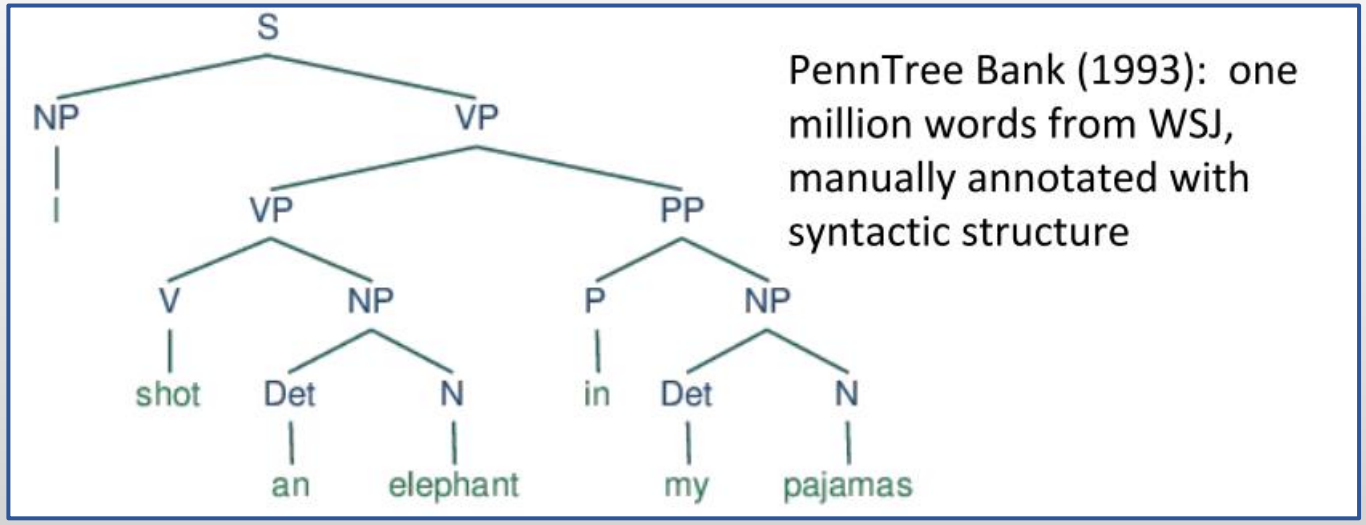
Goal: It aims to place “the” (determiner).

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.



NLP Approaches (cont.)

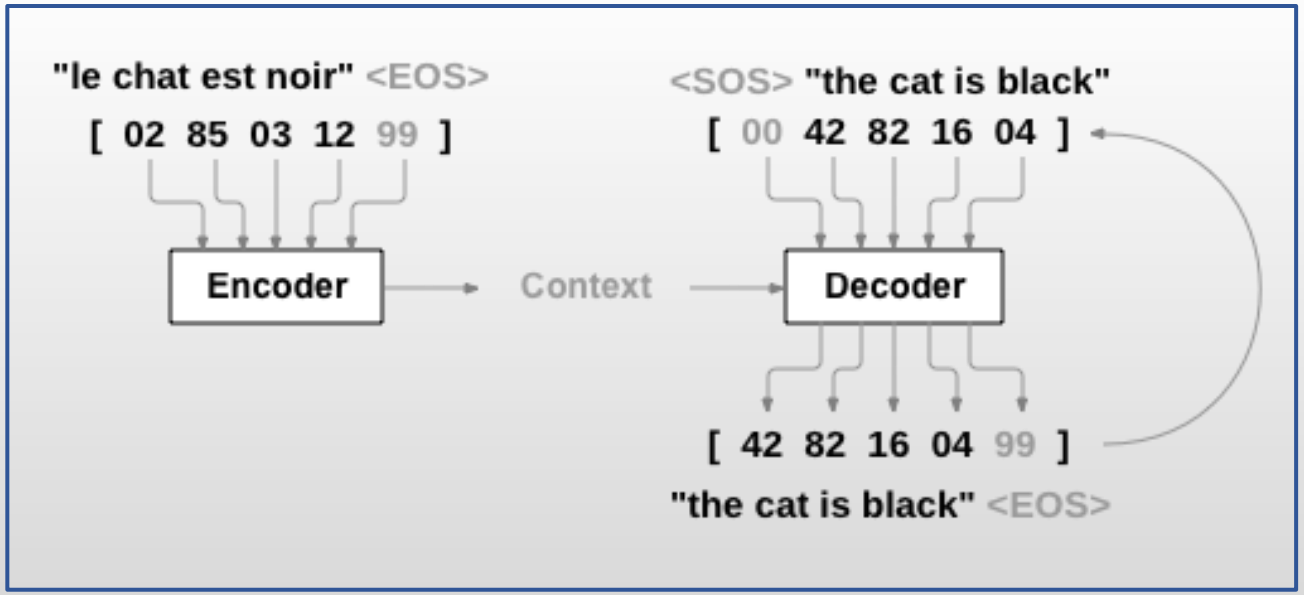
- Statistical or Machine learning approach
  - utilizes language samples to understand the properties of language.
  - In the 1980s, there was a shift towards using probabilistic methods in NLP, which was inspired by information theory.
  - However, this approach has the drawback that it requires the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



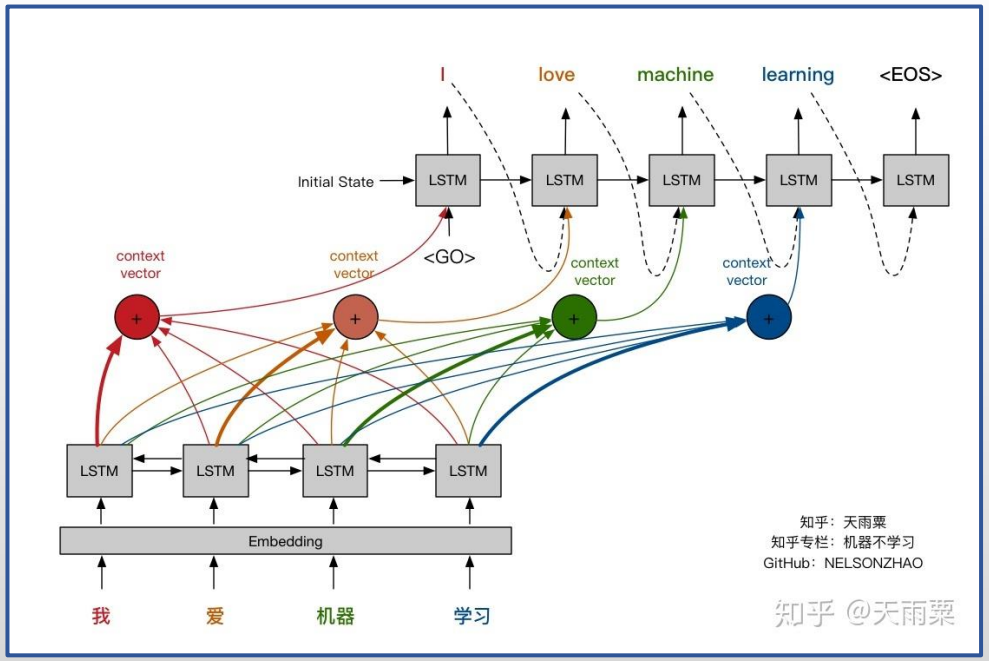
NLP Approaches (cont.)

- 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.



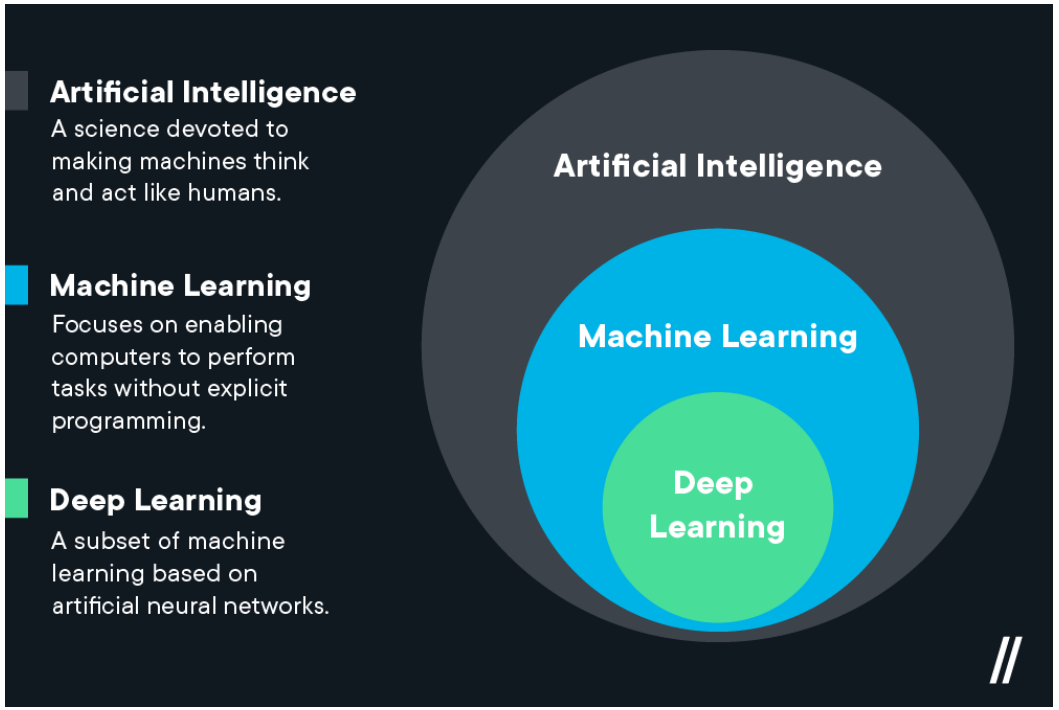
Sequence to sequence model



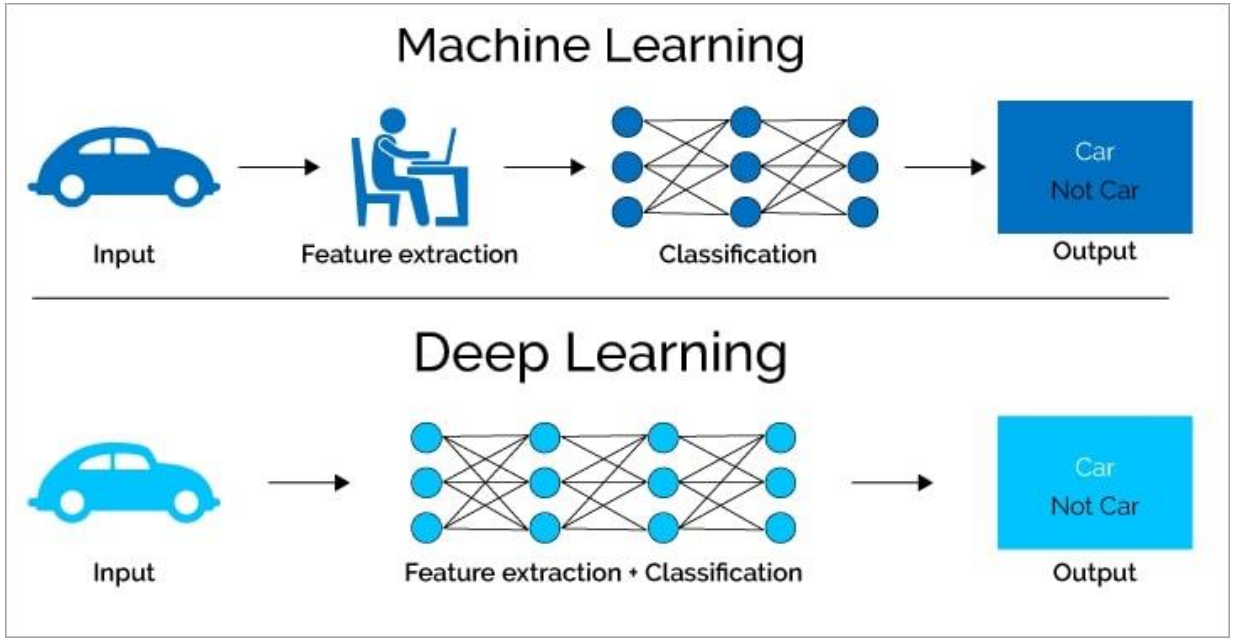
Attention mechanism in sequence to sequence model

Deep Learning in NLP

What is Deep Learning?

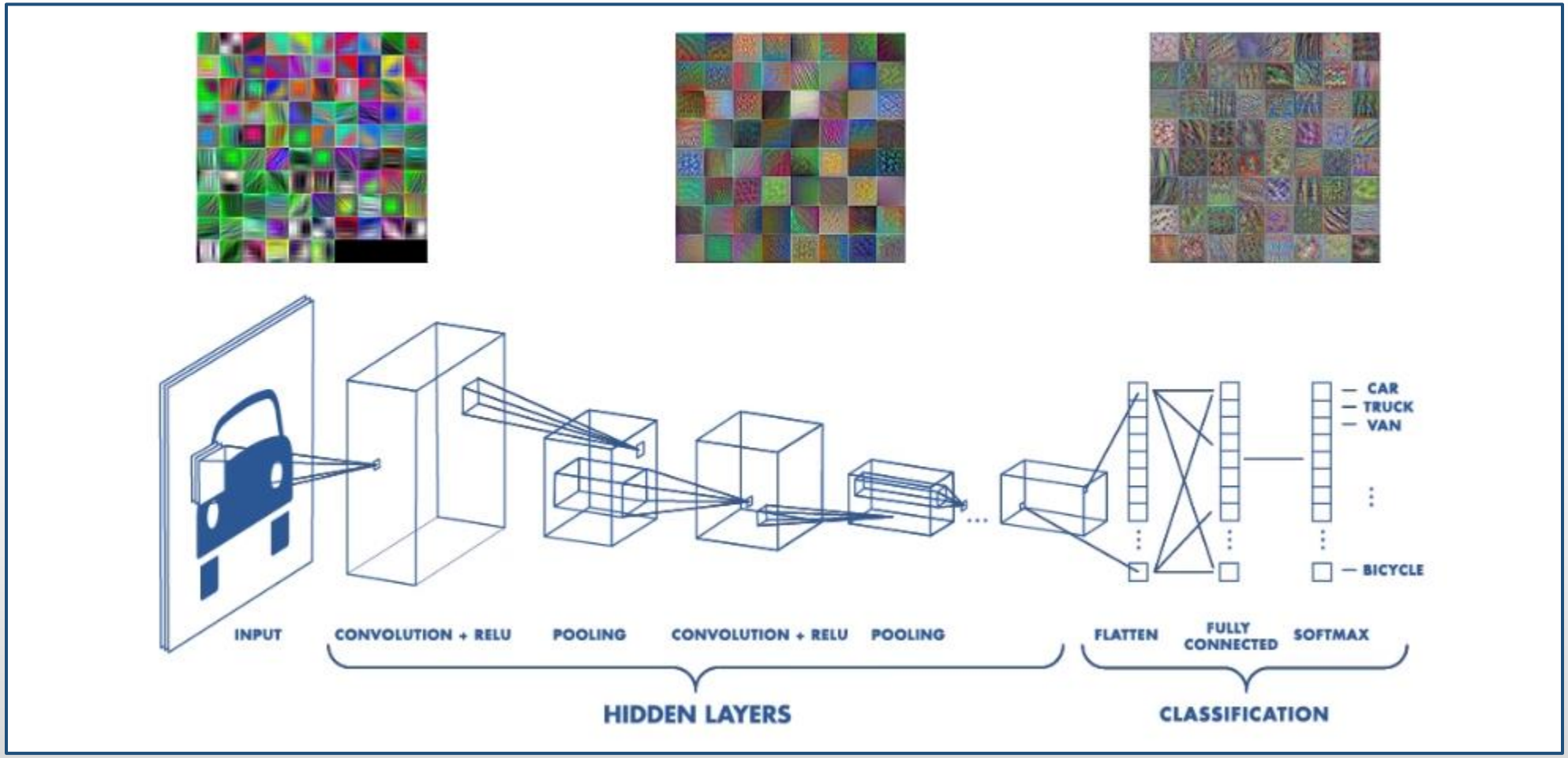


https://flatironschool.com/blog/deep-learning-vs-machine-learning/



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.

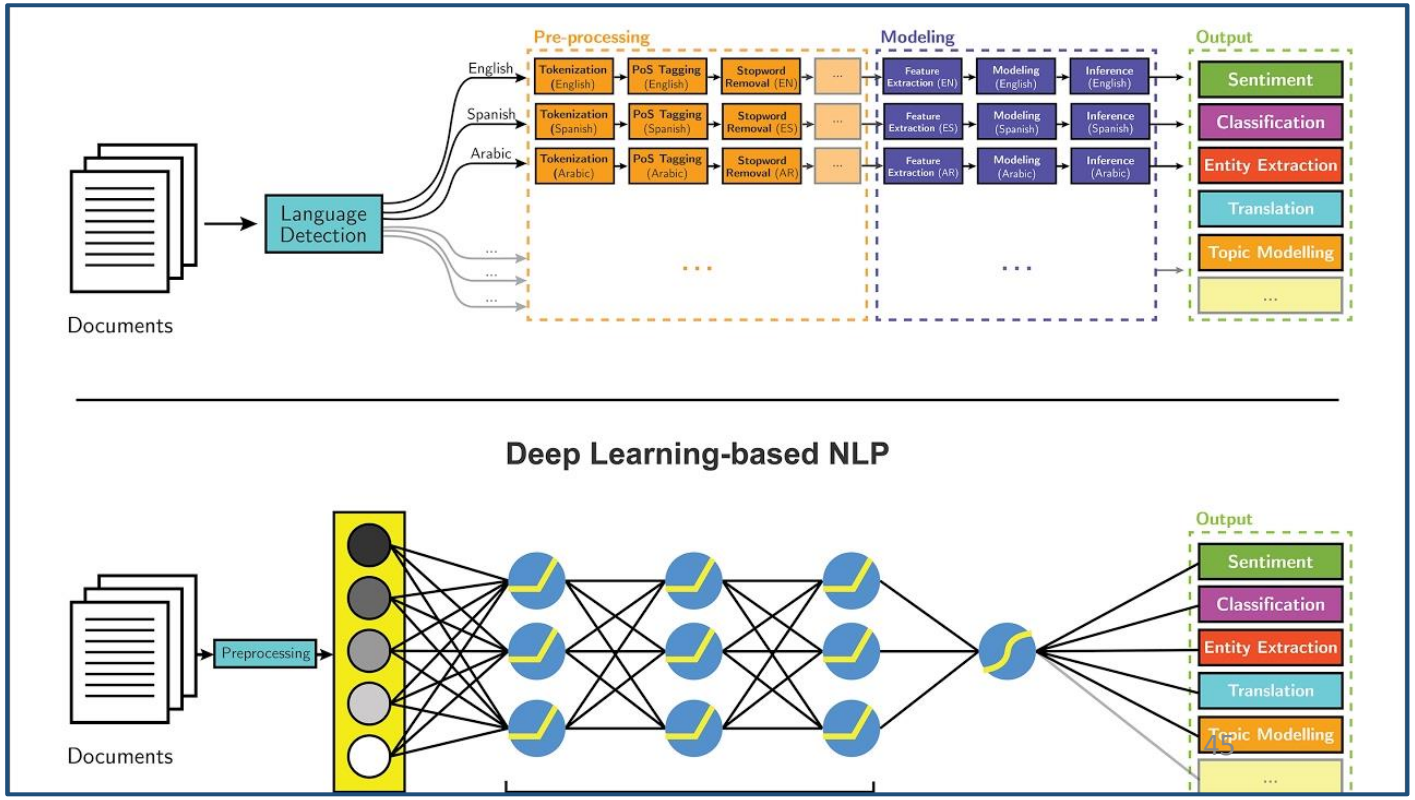
https://www.youtube.com/watch?v=V8qrVleGY5U

[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** Learning models ...
☆ Save 99 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 99 Cite Cited by 27 Related articles All 6 versions »

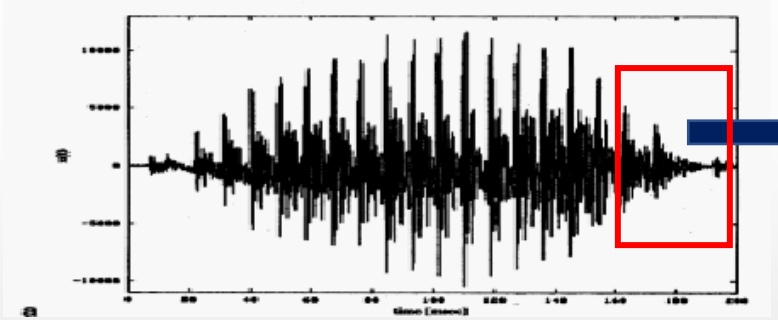
An overview of shallow and **deep natural language processing** for ontology learning
A Zouaq - Ontology learning and knowledge discovery using the ..., 2011 - igi-global.com
... **natural language processing** (**NLP**) techniques ranging from lexical acquisition to shallow and **deep** ... when they come as a complement to the **NLP** process. There are many reasons why ...
☆ Save 99 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 99 Cite Cited by 8 Related articles All 2 versions



Reasons for exploring Deep Learning

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



Speech

$$\begin{bmatrix} .2 \\ .7 \\ .1 \end{bmatrix}$$

images

$$\begin{bmatrix} .5 \\ .1 \\ .6 \end{bmatrix}$$

Words

$$\begin{bmatrix} .1 \\ .5 \\ .8 \end{bmatrix} \begin{bmatrix} .2 \\ .1 \\ .2 \end{bmatrix} \begin{bmatrix} .2 \\ .6 \\ .9 \end{bmatrix} \dots\dots$$

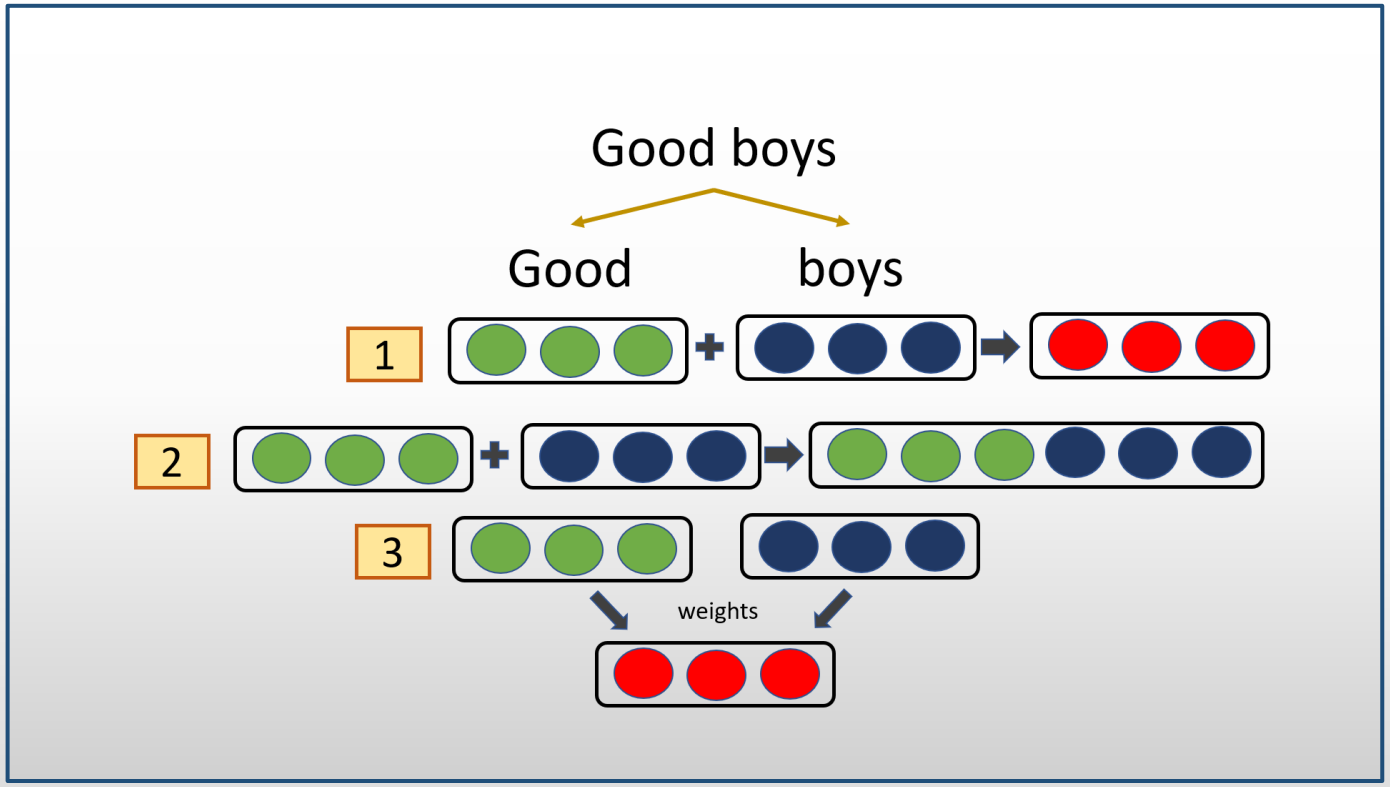
$$\begin{bmatrix} .8 \\ .2 \\ .4 \end{bmatrix}$$

Sentences

Life is not about waiting for the storm to pass. It's about learning to dance in the rain.

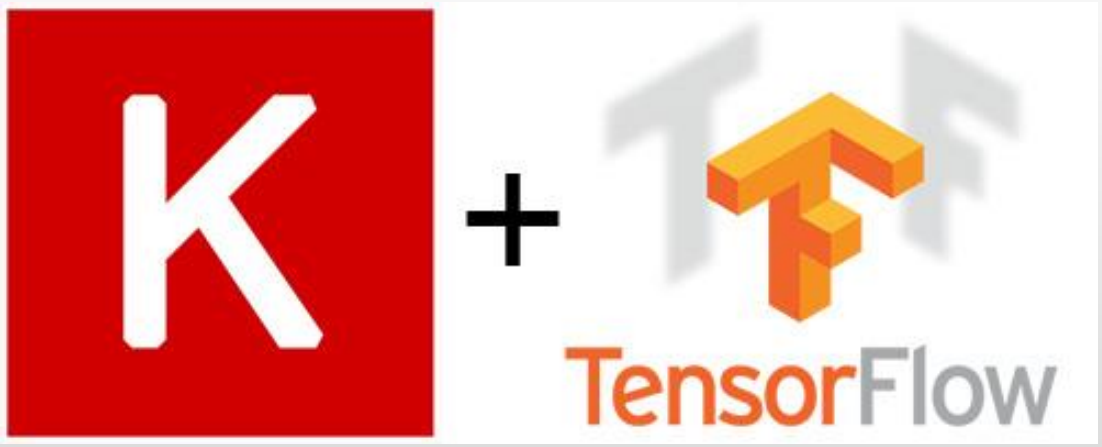
Reasons for exploring Deep Learning (cont.)

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

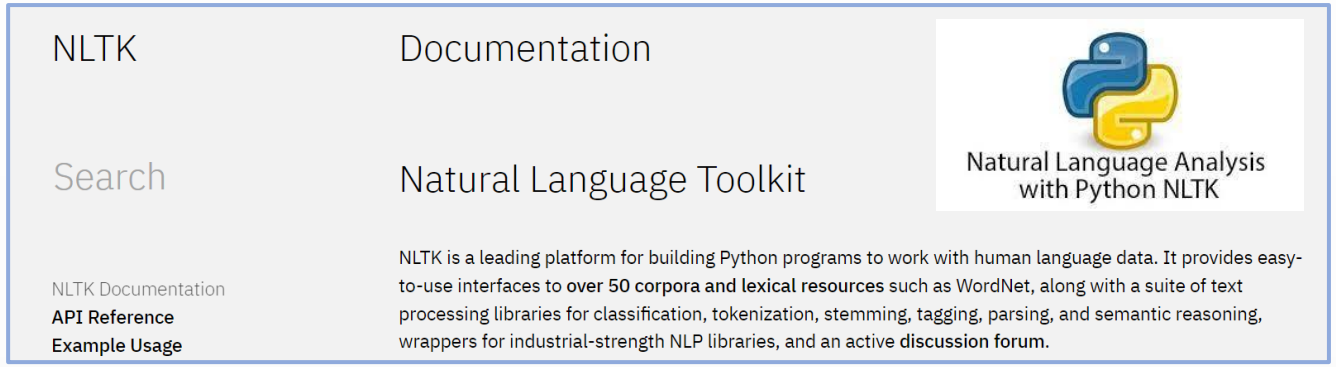


NLP tools

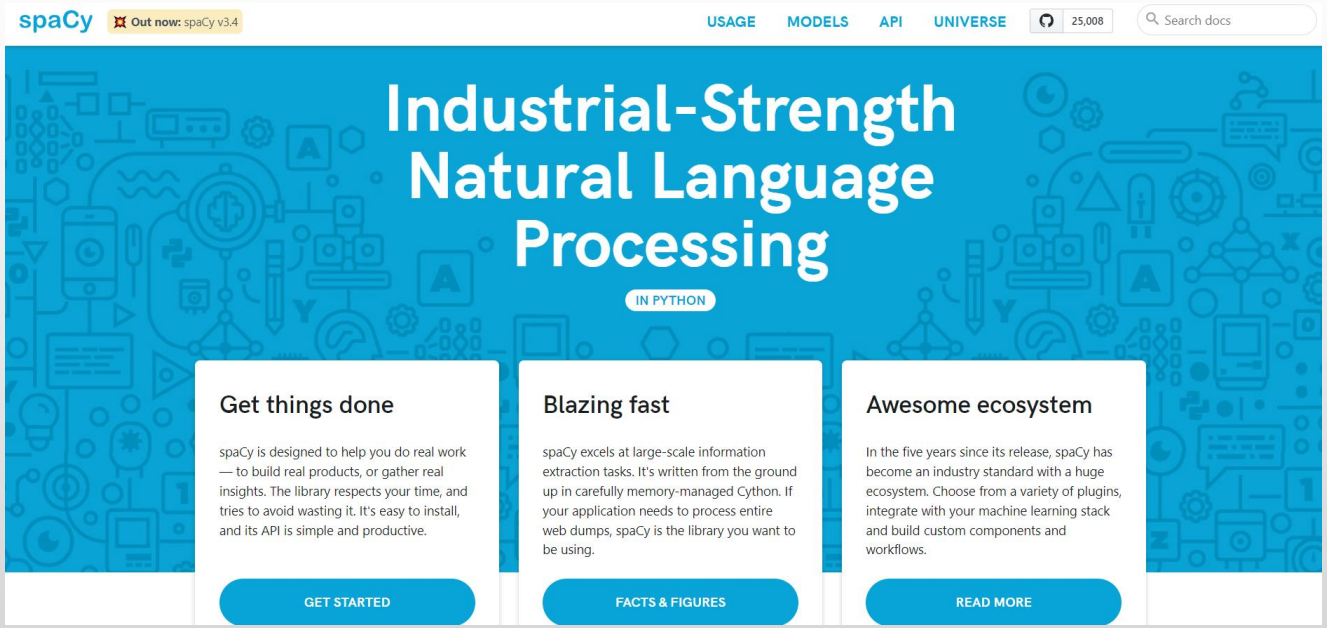
Framework & Implementation



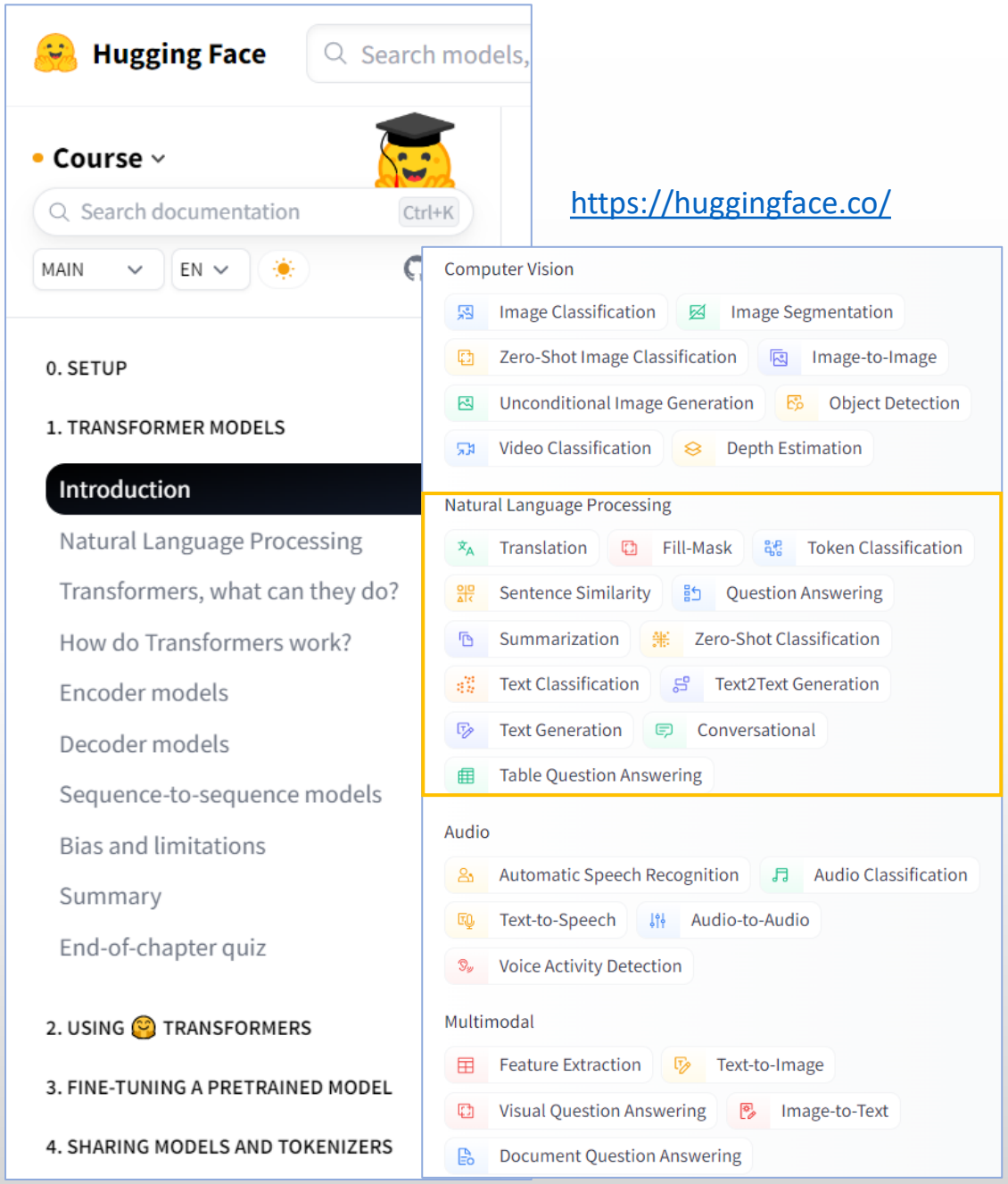
NLP Libraries



https://www.nltk.org/

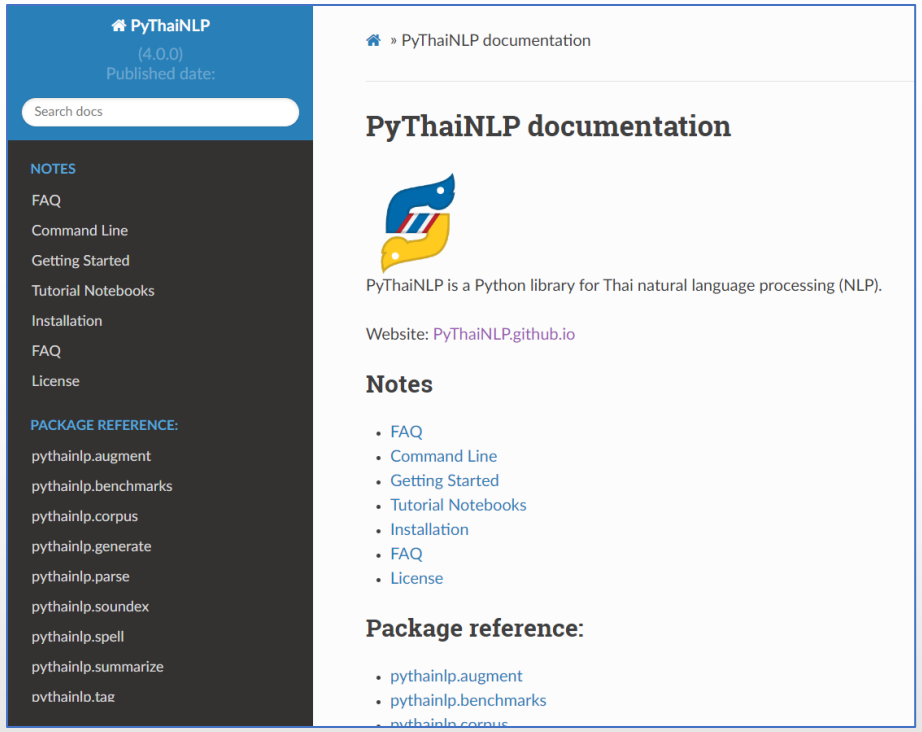


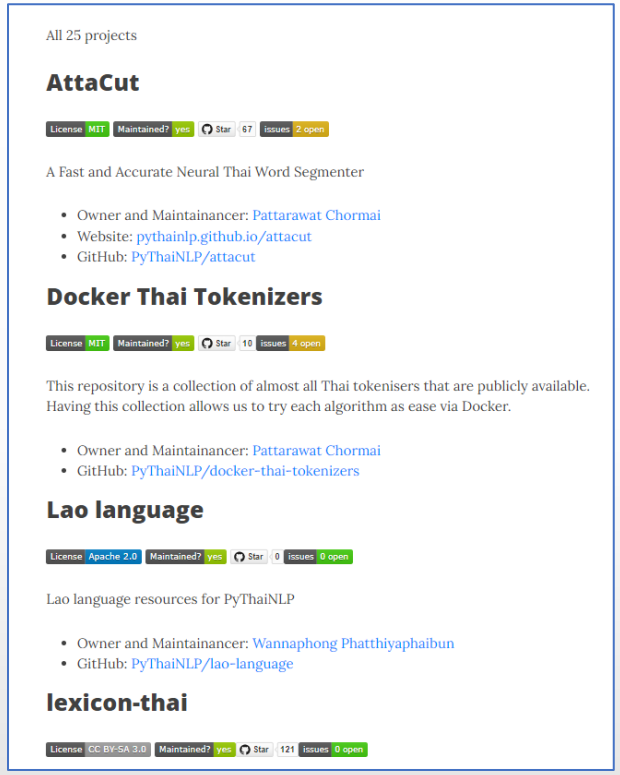
https://spacy.io/



https://huggingface.co/

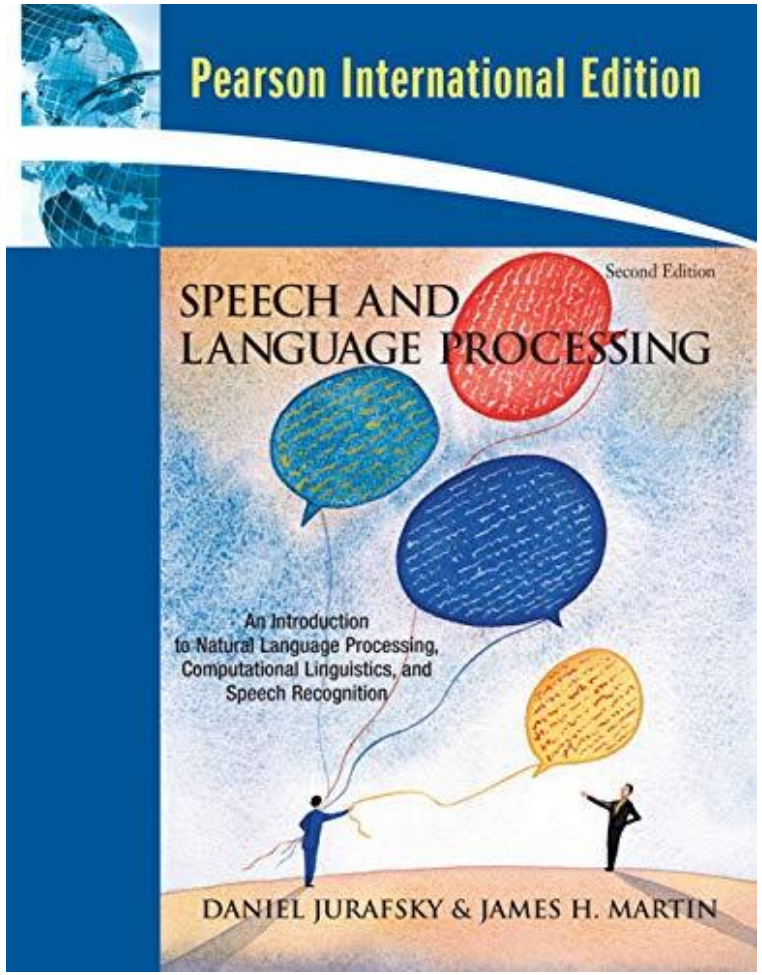
NLP Library for Thai





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