CHULA ΣNGINEERING
Foundation toward Innovation
COMPUTER

Introduction to Natural Language Processing

2110572: NLP SYS

Peerapon Vateekul

Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University

peerapon.v@chula.ac.th

Credits to: TA.Pluem, TA.Knight, and all TA alumni

Outlines

1) What is NLP?

2) History of NLP & Deep Learning

3) NLP System Building Overview & Demo

4) This Course

5) NLP Tools

1) What is NLP?

- Definition & Levels of Understanding
- NLP today

1.1) What is NLP?

Natural Language Processing (NLP)

Technology to handle human language (usually text) using computers

- Aid **human-machine communication** (e.g. question answering, dialog, code generation)
- Aid **human-human communication** (e.g. machine translation, spell checking, assisted writing)

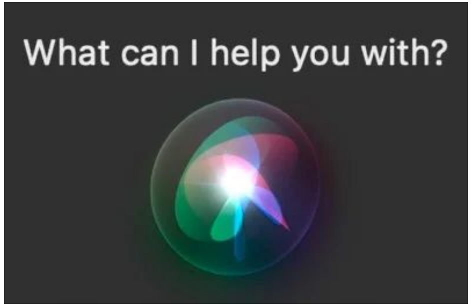
Language (Text, Speech) → **Meaning**

NLP (Interpretation)

I want to eat Japanese food.

May I order Yayoi for you?

AI/ML (Generation)



Going beyond string matching

Goal: Analyze/understand language (NOT just string matching!)

- Syntactic structures, Text classification, Entity/relation linking

We use NLP many times a day without knowing it!

Level of understanding in NLP

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

Lexical Analysis:

Text ⟶ Paragraphs, Sentences, and Words

Syntactic Analysis (Parsing):

Grammar/Relationship between words

Semantic Analysis:

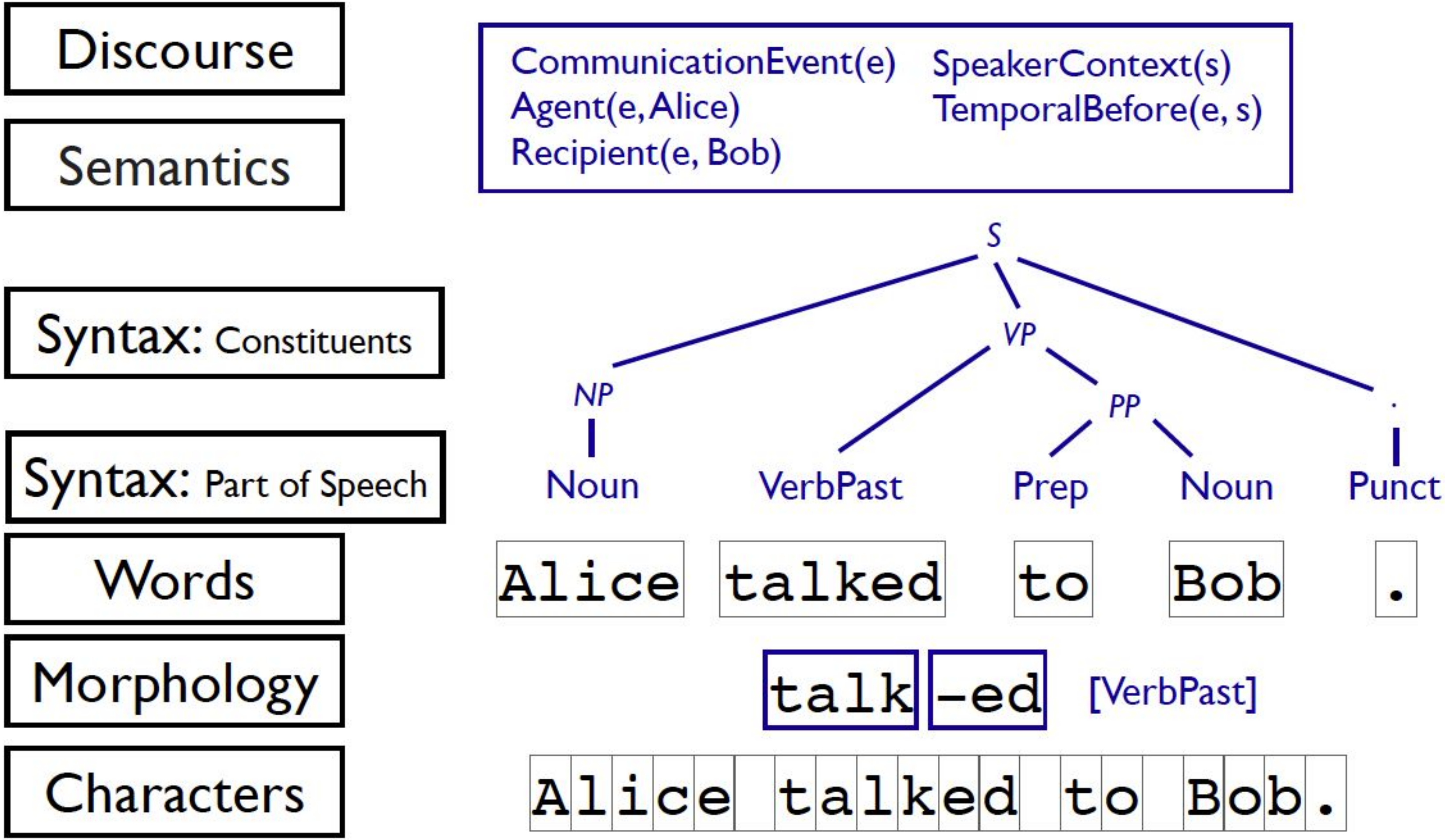
Exact meaning of the sentence

Discourse Integration:

Meaning of the sentence **based on the previous sentence (pronouns)**

Pragmatic Analysis:

Actual Meaning based on **the context** and real-world knowledge


Discourse
Semantics
Syntax: Constituents
Syntax: Part of Speech
Words
Morphology
Characters
CommunicationEvent(e)
Agent(e, Alice)
Recipient(e, Bob)
SpeakerContext(s)
TemporalBefore(e, s)
S
NP
VP
PP
.
Noun
VerbPast
Prep
Noun
Punct
Alice
talked
to
Bob
.
talk
-ed
[VerbPast]
Alice talked to Bob.

Tokenization

- Input: Mr.Smith goes to Washington
- Output: [Mr.Smith, goes, to, Washington]

Part of Speech tagging

- Input: [Mr.Smith,goes,to,Washington]
- Output:[(Mr.Smith,**NNP**), (goes,**VBZ**), (to,**TO**), (Washington,**NNP**)]

PENN Part Of Speech Tags

- NNP – proper noun
- VBZ - Verb, 3rd person singular present
- TO –to

Ref: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

NER

- Input:[(Mr.Smith,**NNP**), (goes,**VBZ**), (to,**TO**), (Washington,**NNP**)]
- Output:[(Mr.Smith,NNP,**PER**), (goes,VBZ,**O**), (to,TO,**O**), (Washington,NNP,**LOC**)]

Named Entity Tags

- PER –Person
- LOC – Location
- ORG – Organization
- O – Other

Application

- e.g. Word Cloud (Named Entity Only)

Tokenization

- Input:ขสมก.เล็งจัดหารถ
- Output: ขสมก.,เล็ง,จัดหา,รถ

Part of Speech tagging

- Input:[ขสมก.,เล็ง,จัดหา,รถ]
- Output:[(ขสมก.,NR),(เล็ง,VV),(จัดหา,VV),(รถ,NN)]

PENN Part Of Speech Tags

- NR – proper noun
- VV - Main verbs in clauses, verb-form
- NN – Non-proper noun

Ref: BEST2010 dataset

NER

- Input:[(ขสมก.,NR),(เล็ง,VV),(จัดหา,VV),(รถ,NN)]
- Output:[(ขสมก.,NR,ORG),(เล็ง,VV,O),(จัดหา,VV,O),(รถ,NN,O)]

Named Entity Tags

- PER –Person
- LOC – Location
- ORG – Organization
- O – Other

Application

- e.g. Word Cloud (Named Entity Only)

1.2) NLP Applications

NLP can Answer Questions

1

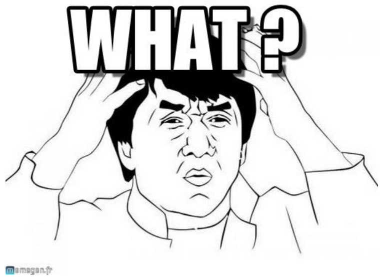
Who is Peerapon Vateekul?

Asst. Prof. Dr. Peerapon Vateekul is a faculty member in the Department of Computer Engineering at Chulalongkorn University in Bangkok, Thailand. CHULALONGKORN UNIVERSITY He earned his Ph.D. in Electrical and Computer Engineering from the University of Miami in 2012. CHULALONGKORN UNIVERSITY His research interests include data mining, machine learning, data warehousing, and business intelligence. CHULALONGKORN UNIVERSITY He has published numerous papers in these fields and is involved in various academic and research activities. DBLP

Sources

Retrieved Dec 31, 2024

...but (sometimes) makes up facts



2

What's special about Peerapon Vatekul's choice in fashion, specifically his socks?

Asst. Prof. Dr. Peerapon Vateekul has been involved in research utilizing vibrating socks to improve gait in patients with Parkinson's disease. These specialized socks provide tactile biofeedback to alleviate freezing of gait, a common symptom in Parkinson's patients. The use of vibrating socks offers the advantage of being discreet, as they are not noticeable to bystanders. RESEARCHGATE

This innovative approach reflects Dr. Vateekul's commitment to applying machine learning and data mining techniques to develop real-world solutions, particularly in the medical field. By integrating technology into wearable items like socks, his research aims to enhance the quality of life for individuals with movement disorders.

3

What makes Peerapon Vateekul socks a timeless fashion? Be brief.

Peerapon Vateekul's vibrating socks are considered timeless in fashion due to their combination of **innovative technology** and **modern design**. They blend functional health benefits with stylish aesthetics, making them not only practical for improving mobility but also a fashionable accessory. Their unique blend of **comfort, utility, and sleek design** allows them to stand out as a forward-thinking product, appealing to both fashion-conscious and health-focused consumers.

Vibratory and plantar pressure stimulation: Steps to improve freezing of gait in Parkinson's disease

October 2022 · Parkinsonism & Related Disorders 105(12)
DOI:10.1016/j.parkreldis.2022.10.024

Authors:

Warongporn Phuenpathom, Chulalongkorn University
Pattamon Panyakaew, Chulalongkorn University
Peerapon Vateekul, Chulalongkorn University

Retrieved Dec 31, 2024 https://www.researchgate.net/publication/309344097_Alleviating_Freezing_of_Gait_using_phase-dependent_tac

NLP can Translate Text

Google

google translate

All Images Maps News Videos More Settings Tools

About 1,180,000,000 results (0.39 seconds)

English	Thai
As the new year gets underway, expert commentators give their view on what 2018 holds in store.	เป็นปีใหม่ที่กำลังได้รับการแสดงความคิดเห็นของผู้เชี่ยวชาญให้มุมมองของพวกเขาเกี่ยวกับสิ่งที่ 2018 เก็บในร้าน
Here are three big themes to watch out for over the next 12 months.	ต่อไปนี้เป็นหัวข้อใหญ่สามข้อที่ควรระวังในช่วง 12 เดือนข้างหน้า
Can the stock market rally go on? The new year has begun with stock markets in the UK and US hitting new record highs.	การชุมนุมตลาดหุ้นสามารถดำเนินต่อไปได้หรือไม่? ปีใหม่เริ่มมีตลาดหุ้นในสหราชอาณาจักรและสหรัฐฯพุ่งสูงเป็นประวัติการณ์
The Dow Jones Industrial Average rose above 25,000 points for the first time this week, while the broader S&P 500 is also at historic highs.	ดัชนีเฉลี่ยอุตสาหกรรมดาวโจนส์ปรับตัวสูงขึ้นกว่า 25,000 จุดเป็นครั้งแรกในสัปดาห์นี้ ขณะที่ดัชนี S & P 500 ที่ใหญ่ขึ้นก็อยู่ในระดับสูงเป็นประวัติการณ์

Markets, Brexit and Bitcoin: 2018's themes

By Chris Johnston
Business reporter

5 January 2018

Share

As the new year gets underway, expert commentators give their view on what 2018 holds in store.

http://www.bbc.com/news/business-42581934

...but (sometimes) loses Translation Meaning

Detect language | **Thai** | English | Spanish → Thai | **English** | Chinese (Simplified)

แล้วเธอจะเจอดี

Sarcasm

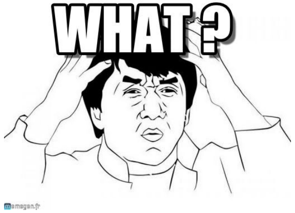
Then you will meet with good things.

ลำตัดคณะนี้เล่นถึงพริกถึงขิงจริง

Uncommon Words

Idioms

This group cut the faculty to play to the real ginger.

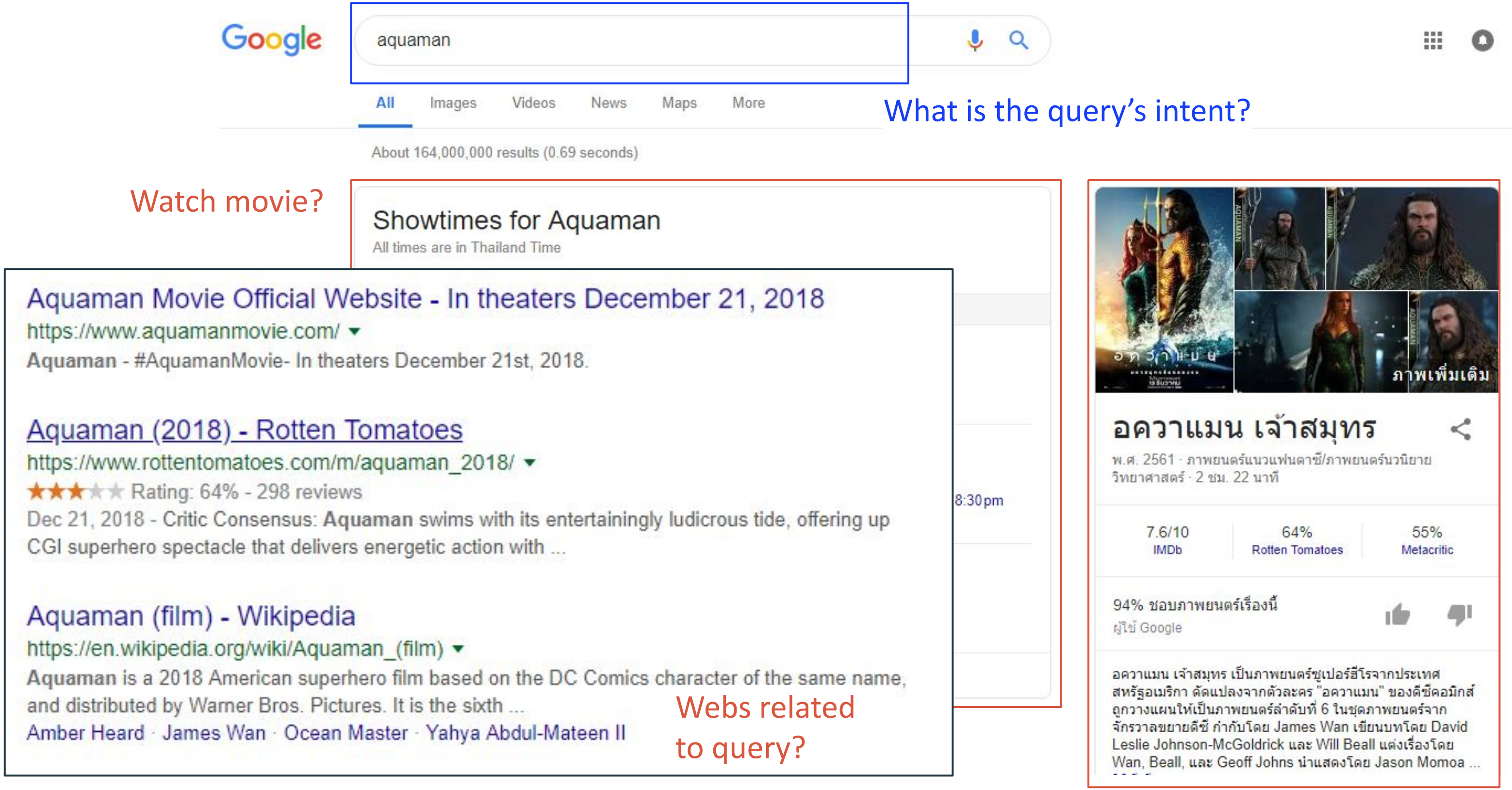


Retrieved Dec 31, 2024

NLP asks “why” are we Searching

What is the query’s intent?

Watch movie?



Webs related to query?

Review/ Summarize?

NLP can Extract Information from Text

Data science perspective on clinical research

Abstract clinical records 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		

Pathology Report: REMOVED_ACCESSION_ID
ACCESSIONED ON: REMOVED_DATE
CLINICAL DATA: Carcinoma **right breast**.
*** FINAL DIAGNOSIS ***
LYMPH NODE (SENTINEL), EXCISION
(REMOVED_CASE_ID): METASTATIC
CARCINOMA IN 1 OF 1 LYMPH NODE.
NOTE: The metastatic deposit spans 0.19cm and is identified on H&E and cytokeratin immunostains. A second cytokeratin-positive but cauterized focus likely also represents metastatic tumor (<0.1cm). There is **no evidence of extranodal extension**.
BREAST (RIGHT), EXCISIONAL BIOPSY
(REMOVED_ACCESSION_ID :
REMOVED_CASE_ID -B): **INVASIVE DUCTAL CARCINOMA** (SEE TABLE #1). **DUCTAL CARCINOMA IN-SITU**, GRADE 1. **ATYPICAL DUCTAL HYPERPLASIA**. **LOBULAR NEOPLASIA (ATYPICAL LOBULAR HYPERPLASIA)**.
TABLE OF PATHOLOGICAL FINDINGS #1

Name	Extraction
Breast Side	Right
Ductal Carcinoma in Situ	Present
Invasive Lobular Carcinoma	Absent
Invasive Ductal Carcinoma	Present
Cancer	Present
Lobular Carcinoma in Situ	Absent
Atypical Ductal Hyperplasia	Present
Atypical Lobular Hyperplasia	Present
Lobular Neoplasia	Present
Flat Epithelial Atypia	Absent
Blunt Adenosis	Absent
Atypia	Present
Positive Lymph Nodes	Present
Extracapsular Axillary Nodal Extension	Absent
Isolated Cancer Cells in Lymph Nodes	Absent
Lymphovascular Invasion	Absent
Blood Vessel Invasion	Absent
Estrogen Receptor Status	Positive
Progesterone Receptor Status	Positive
HER 2 (FISH) Status	Unknown

Parsing pathology reports into database

...but sometimes fail at Basic Tasks

From Bangkok Post Moo Deng story 31 Dec. 2024

NER by Stanford CoreNLP

PERSON TITLE PERSON PERSON DATE 2025
Suan Dusit Poll president Pornpan Buathong said the popularity of the Moo Deng story , which captivated much of the world , reflected the sentiment felt by Thais this year , made stressful by hard

ORGANIZATION
news like the Icon Group case and the end of Move Forward .

Different tools can be failed in various cases:

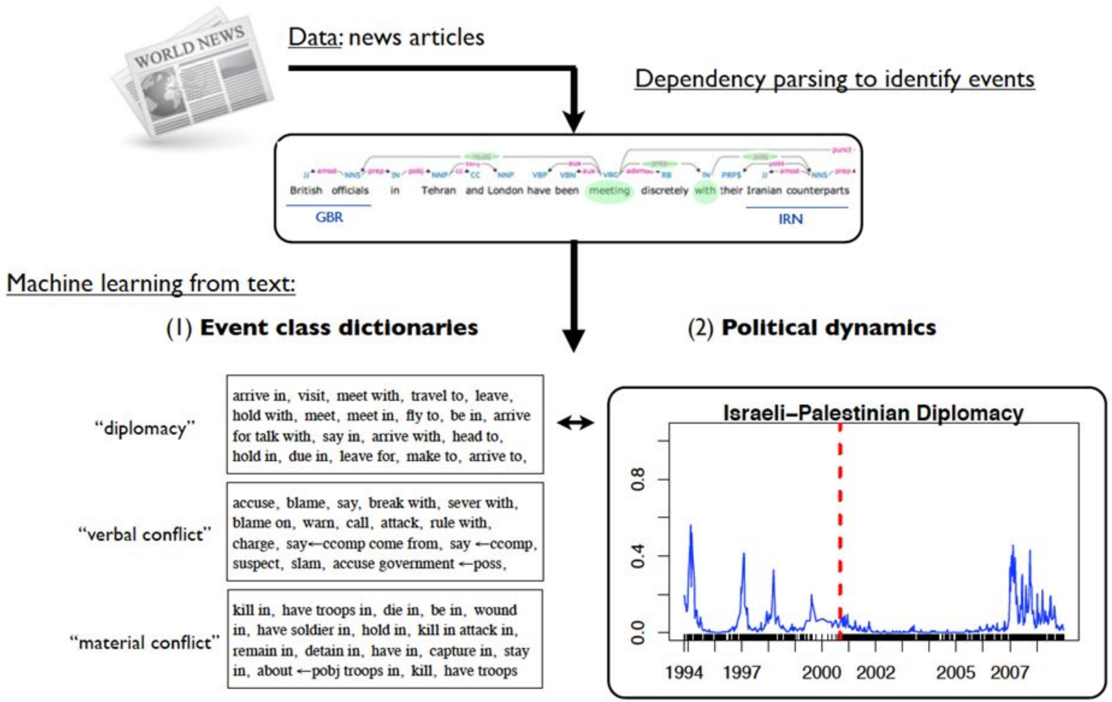
- spaCy: CANNOT capture "president" (title)
- Stanford: CANNOT capture "Move Forward" (ORG)
- Both fail: "Suan Dusit" as PERSON

NER by spaCy

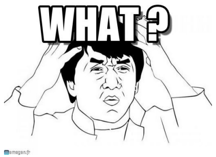
Suan Dusit PERSON Poll president Pornpan Buathong PERSON said the popularity of the Moo Deng PERSON story, which captivated much of the world, reflected the sentiment felt by Thais GPE this year DATE , made stressful by hard news like the Icon Group ORG case and the end of Move Forward ORG .

Bangkok Post - Moo Deng voted best story of 2024, beats Icon Group fraud 31 Dec 2024
https://www.bangkokpost.com/thailand/general/2930405/moo-deng-voted-best-story-of-2024-beats-icon-group-fraud

NLP can Analyze Trends



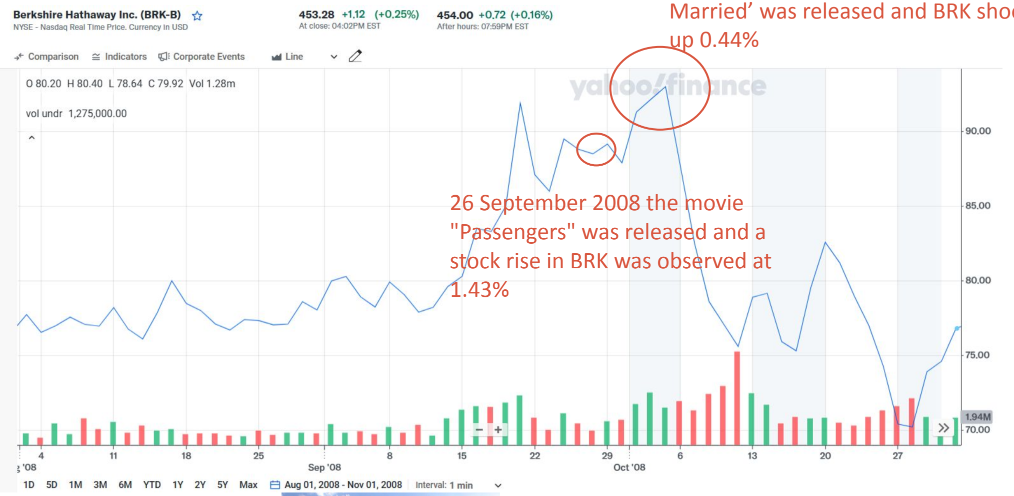
...but (sometimes) fails on random things



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."

October 3, 2008, 'Rachel Getting Married' was released and BRK shoot up 0.44%



26 September 2008 the movie "Passengers" was released and a stock rise in BRK was observed at 1.43%


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

- A Message From Warren E. Buffett
- Annual & Interim Reports
- Special Letters From Warren & Charlie RE:Past, Present and Future
- Link to SEC Filings
- Links to Berkshire Subsidiary Companies
- Corporate Governance
- News Releases
- Warren Buffett's Letters
- Charlie Munger's Letter
- Annual Meeting Inform
- Celebrating 50 Years of
- Comparative Rights and

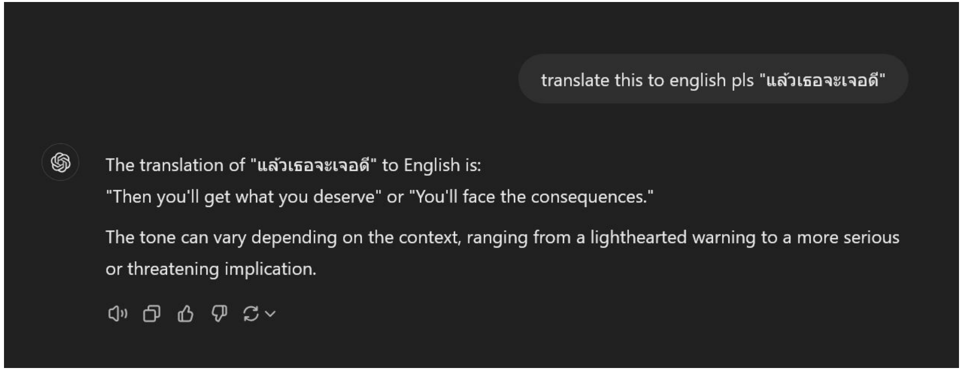
NLP is difficult!

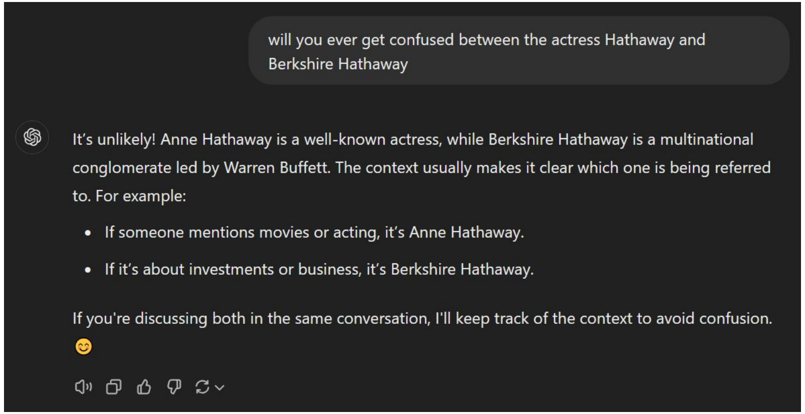
Word-level ambiguity!

Ref: Prof. Regina Barzilay, NLP @MIT

Many problems are trivialized in the LLM Era

In November 2022, OpenAI released ChatGPT and thus began the LLM era.





2) History of NLP & Deep Learning

- History of NLP
- Deep Learning

A Brief Timeline of NLP

Logical Era — 1950

- Alan Turing's Turing Test
- Chomsky's 'universal grammar'[1]
- Interest in automatic translation[2]
- MIT ELIZA rule-based rephrasing
- MIT SHRDLU moving blocks
- Conceptual ontologies to chatterbots

Statistical Era — 1980

- Decision trees (hard rule-based)
- Statistical models
- Emergence of large textual corpora
- 1993 IBM alignment models for statistical machine translation

Neural Era — 1990

- 1990 Elman network word embedding
- Recurrent Neural Networks (RNN)
- 1995 Improved RNN as LSTMs
- 1997 Bidirectional recurrent neural networks (BRNN)
- 2006 Bi-LSTM in speech recognition and text-to-speech

Neural (DL) Era — 2010

- 2012 ImageNet Alexnet
- 2013 Word2Vec and GloVe
- 2015 Google BERT
- Attention and Transformers
- Pre-trained Models and Transfer Learning (ULMFiT)
- 2019 OpenAI GPT-2

Bag of words ⟶ Word Embeddings ⟶ RNN ⟶ LSTM ⟶ Bi-LSTM ⟶ Attention ⟶ Transformers

[1] Noam Chomsky's Syntactic Structures, a rule-based system of syntactic structures
[2] Georgetown experiment in 1954 to translate 60 Russian sentences into English
https://en.wikipedia.org/wiki/History_of_natural_language_processing

A Brief Timeline of NLP since 2020 (LLM)

By company as of 2023: OpenAI, Google, Meta

2020

- GPT-3 and Large Language Models

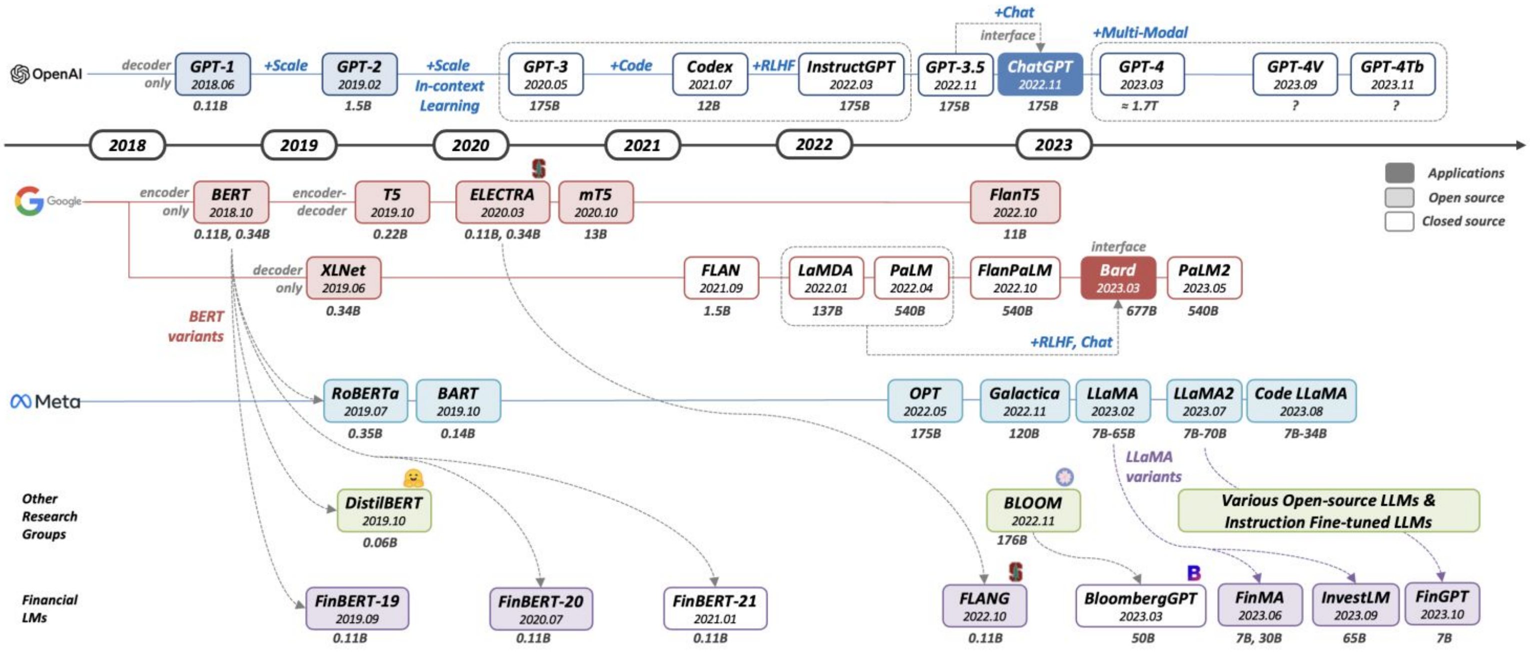


Figure 1: Timeline showing the evolution of selected PLM/LLM releases from the general domain to the financial domain.

2024 Lee et. al. A Survey of Large Language Models in Finance (FinLLMs)
https://arxiv.org/abs/2402.02315

A Brief Timeline of NLP since 2020 (LLM)

Based on timeline as of 2023 (overall companies)

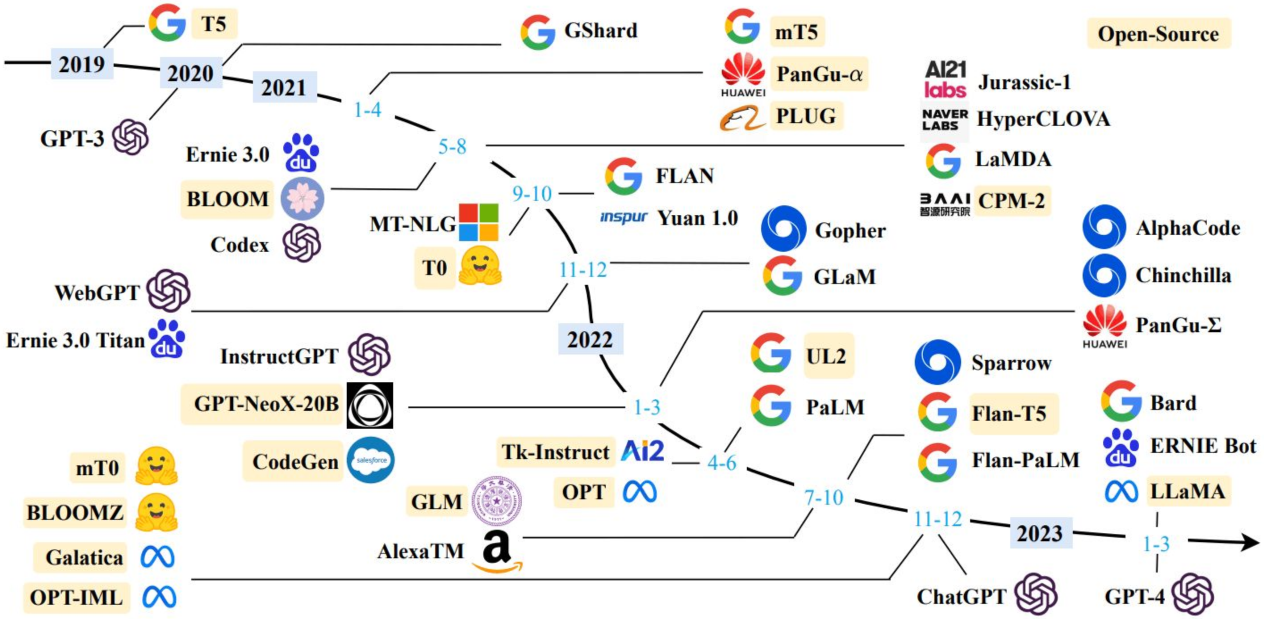


Fig. 1. A timeline of existing large language models (having a size larger than 10B) in recent years. We mark the open-source LLMs in yellow color.

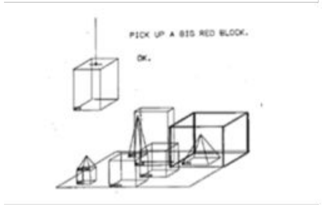
A timeline of existing large language models in recent years.
2023 Zhao et. al. A Survey of Large Language Models https://arxiv.org/abs/2303.18223 [Cited by 2639]

History: Logical Era

1) Symbolic approach

Noam Chomsky, MIT

- Encode all the required information into computer
- In the 1960s and 1970s, Noam Chomsky (an eminent linguist) believed that statistical techniques would never be sufficient to gain a deep understanding of human language.
- This led to the dominance of **knowledge-based approaches**, requiring **human experts** to encode knowledge into computers.
- Disadvantage: It is required substantial human effort.



The dialog that was used as a SHRDLU demo:

Person: PICK UP A BIG RED BLOCK.
Computer: OK. (does it)
Person: GRASP THE PYRAMID.
Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
Person: 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. (does it)
Person: WHAT DOES THE BOX CONTAIN?
Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.
Person: WHAT IS THE PYRAMID SUPPORTED BY?
Computer: THE BOX.
Person: HOW MANY BLOCKS ARE NOT IN THE BOX?

```
(DEFTHEOREM TC-GRASP
        (THCONSE (X Y)   (#GRASP $?X)
                 (THGOAL(#MANIP $?X))
                 (THCOND ((THGOAL (#GRASPING $?X)))
                         ((THGOAL (#GRASPING $_Y))
                          (THGOAL (#GET-RID-OF $?Y)
                                  (THUSE TC-GET-RID-OF))))
                         (T))
                 (THGOAL (#CLEARTOP $?X) (THUSE TC-CLEARTOP))
                 (THSETQ $_Y (TOPCENTER $?X))
                 (THGOAL (#MOVEHAND $?Y)
                         (THUSE TC-MOVEHAND))
                 (THASSERT (#GRASPING $?X))))

(DEFTHEOREM TC-PUT
        (THCONSE (X Y Z)  (#PUT $?X $?Y)
                 (CLEAR $?Y (SIZE $?X) $?X)
                 (SUPPORT $?Y (SIZE $?X) $?X)
                 (THGOAL (#GRASP $?X) (THUSE TC-GRASP))
                 (THSETQ $_Z (TCENT $?Y (SIZE $?X)))
                 (THGOAL (#MOVEHAND $?Z)  (THUSE TC-MOVEHAND))
                 (THGOAL (#UNGRASP)  (THUSE TC-UNGRASP))))
```

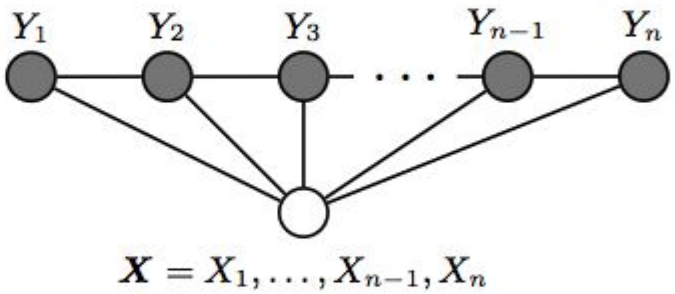
The Internals of SHRDLU

Ref: Prof. Regina Barzilay, NLP @MIT

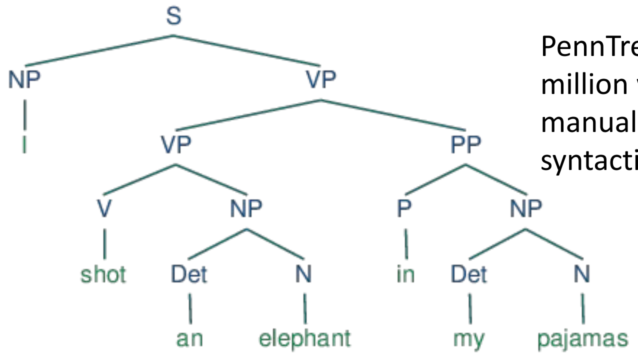
History: Statistical Era

2) Statistical approach

- Infer language properties from language samples
- In 1980s, an empirical revolution took place. Inspired by information theory, it began using **probabilistic approaches** in NLP.
- Disadvantage: It requires handcrafted features.



Conditional Random Fields (CRF)



PennTree Bank (1993): one million words from WSJ, manually annotated with syntactic structure

Ref: Prof. Regina Barzilay, NLP @MIT

Case Study: Determiner placement

Symbolic vs. statistical approaches

Goal: Where 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.

Types of Determiner

Articles	Demonstrative	Possessive Adjectives
the an A	this that these those	my, your his, her its, our your, their
Quantifiers	**Numbers**	**Ordinals**
some, any few, little more, much any, every	one, two three, four twenty, hundred	First, Second Third, Last next

www.links2learn.co.uk

Case Study: *Determiner placement* (cont.)

1) Symbolic approach

- Determiner placement is largely determined by:
  - Type of noun (countable, uncountable)
  - Uniqueness of reference
  - Information value (given, new)
  - Number (singular, plural)

- However, many exceptions and special cases play a role:
  - The definite article is used with newspaper titles (The Times), but zero article in names of magazines and journals (Time)

- Hard to manually encode this information!

Ref: Prof. Regina Barzilay, NLP @MIT

Case Study: Determiner placement (cont.)

2) Statistical approach

- Consider it as classification
- Predictions: {-1, +1}
- Features:
  - Plural?
  - first appearance in text?
  - head token
  - …

"lazy monkeys"

$$\Downarrow$$

$$[1\ 1\ 0\ 0\ 0\ \ldots\ 1]^T$$

-1

"the United States"

$$\Downarrow$$

$$[1\ 1\ 0\ 0\ 0\ \ldots\ 0]^T$$

+1

Minnen et al.	83.58%
Turner&Charniak	86.74%
Knight&Chander	78%

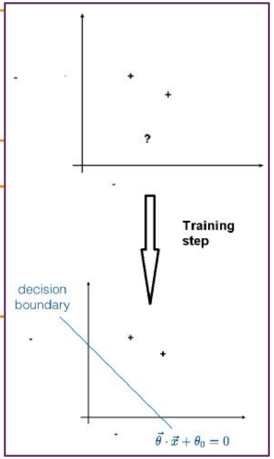
Limitation of traditional statistical approach

1
- Data collection
- Obtain the labels {-1, +1}

2
- Feature Extraction
- The most common features in NLP is word n-grams $\longrightarrow$ encoded as one-hot vector

3
- Training

4
- Testing

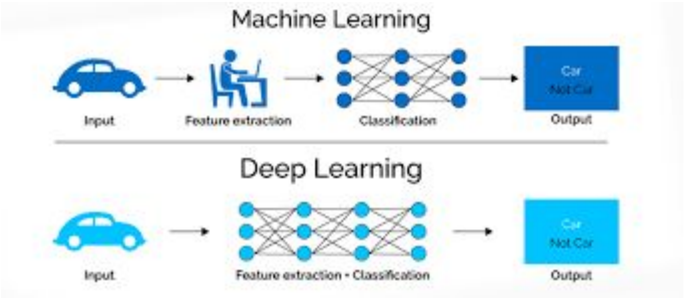


- Sparsity:
  - feature vectors are typically high-dimensional and sparse (i.e., most elements are 0).
- Feature engineering:
  - Need experts to manually design features

Map **discrete**, one-hot vectors into low-dimensional **continuous** representations.

***** Self-learned features $\longrightarrow$ Deep Learning *****

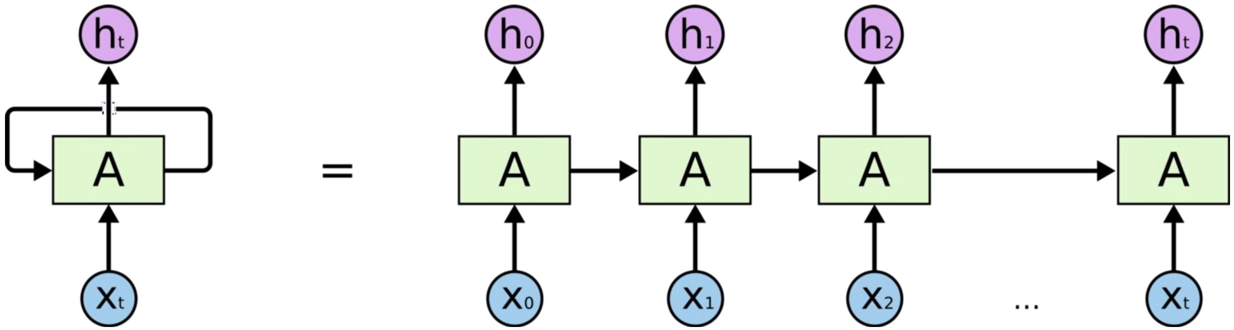
pear	*apple*
$[1\ 0\ 0\ 0\ \ldots\ 0]$	$[0\ 0\ 1\ 0\ \ldots\ 0]$
$\downarrow$	$\downarrow$
$[0.4\ 0.1\ 0.1]$	$[0.6\ 0.2\ 0.3]$



History: Neural Era

3) Deep Learning approach:

- It is a **feature-engineering embedded** neural approach.
- Since the 2010s, it has been gaining a lot of attention and showing many successes.



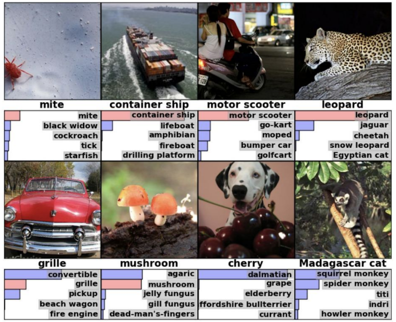
LSTM

Ref: Prof. Regina Barzilay, NLP @MIT

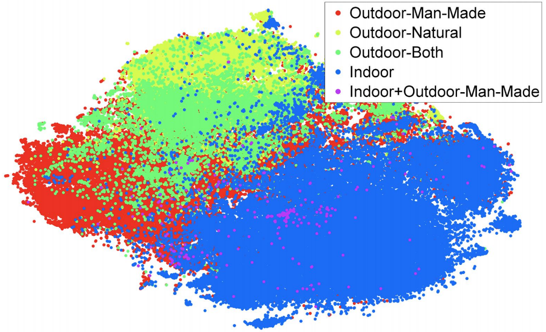
The Spark of DL in Various Domains: Success of AlexNet in Large Scale Visual Recognition Challenge (ILSVRC) 2012

IMAGENET

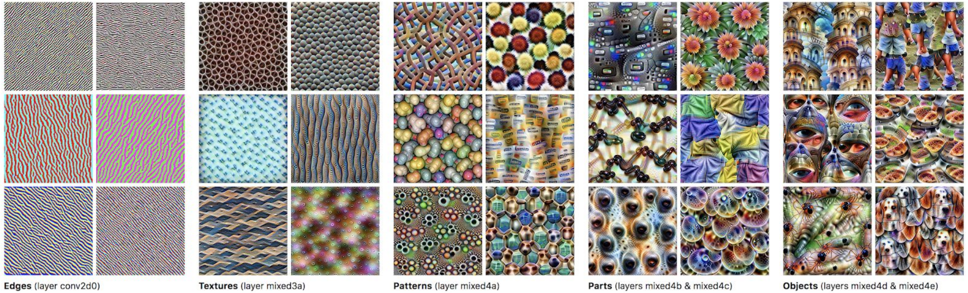
- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.



Deng, J. and Dong, W. and Socher, R. and Li, L.-J. and Li, K. and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

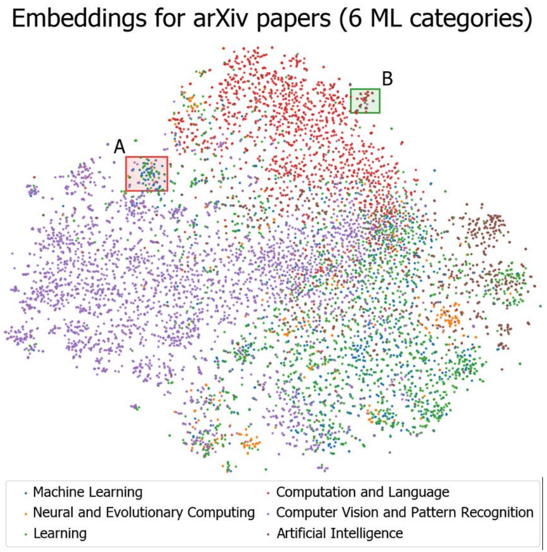


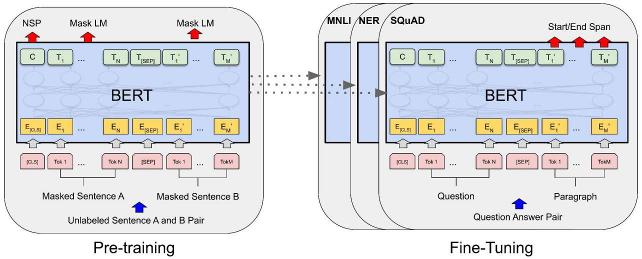
Features trained on ILSVRC-2012 generalize to the SUN-397 dataset. [Donahue et al., 2014]

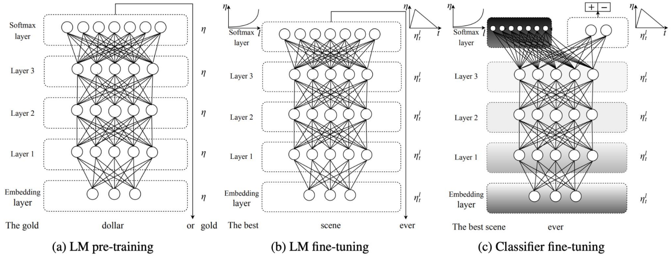


Visualization of the information captured by features across different layers in GoogLeNet trained on ImageNet. (Source: Distill)

NLP “Pretrained” Word Embeddings







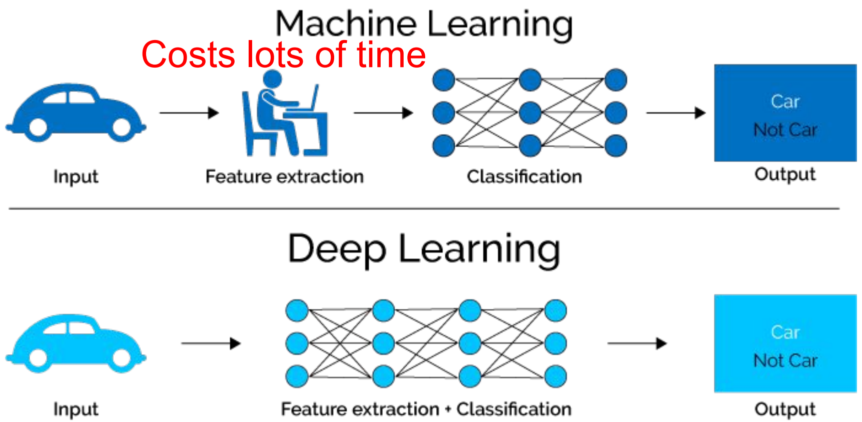
2013 Word2Vec GloVe

2018 Bert ULMFiT

What is Deep Learning?

Part of the machine learning field of learning representations of data. Exceptional effective at learning patterns.

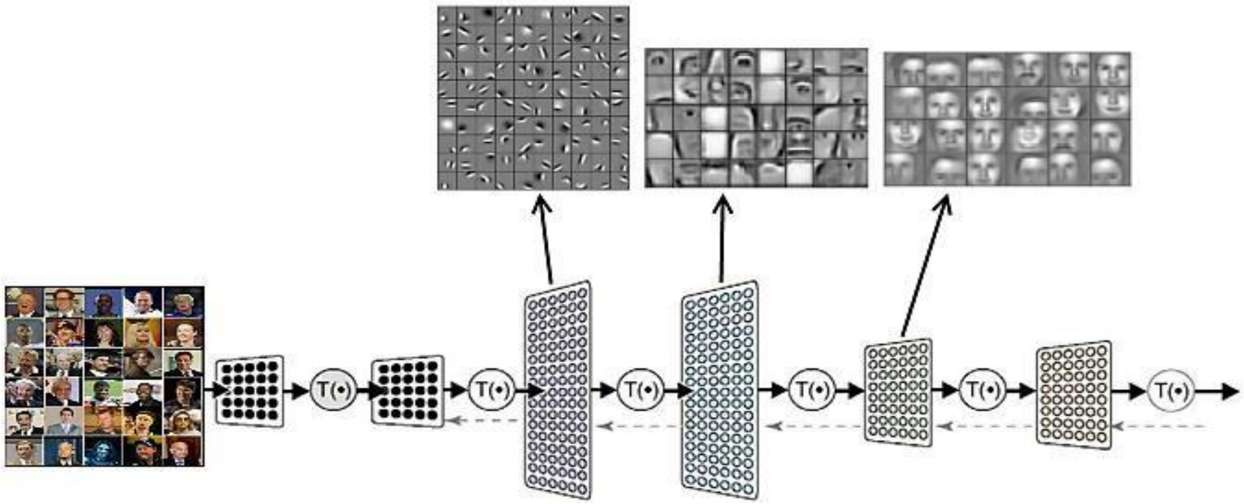
Utilizes learning algorithms that derive meaning out of data by using a hierarchy of multiple layers that mimic the neural networks of our brain.



Deep Learning – Basics (cont.)

What does it learn?

- A deep neural network consists of a hierarchy of layers, whereby each layer transforms the input data into more abstract representations (e.g. edge -> nose -> face).
- The output layer combines those features to make predictions.

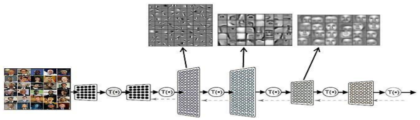


Deep Learning Application

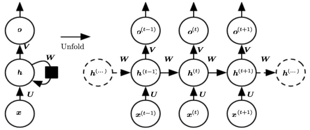
Speech Recognition

Computer Vision

Natural Language Processing



CNN



RNN (LSTM)

NLP + Deep Learning = Deep NLP

- Modern NLP techniques are based on deep learning models.
- These models have obtained very high performance across various NLP tasks.
- They often do not require traditional linguistic feature engineering to perform well.

Deep NLP + LLM (since 2020)

- Multiple orders of magnitude larger than previous generations of deep learning models.
- State-of-the-art performance across a wide range of tasks.
- Usually perform well out of the box. No training is required.

CS324 - Large Language Models

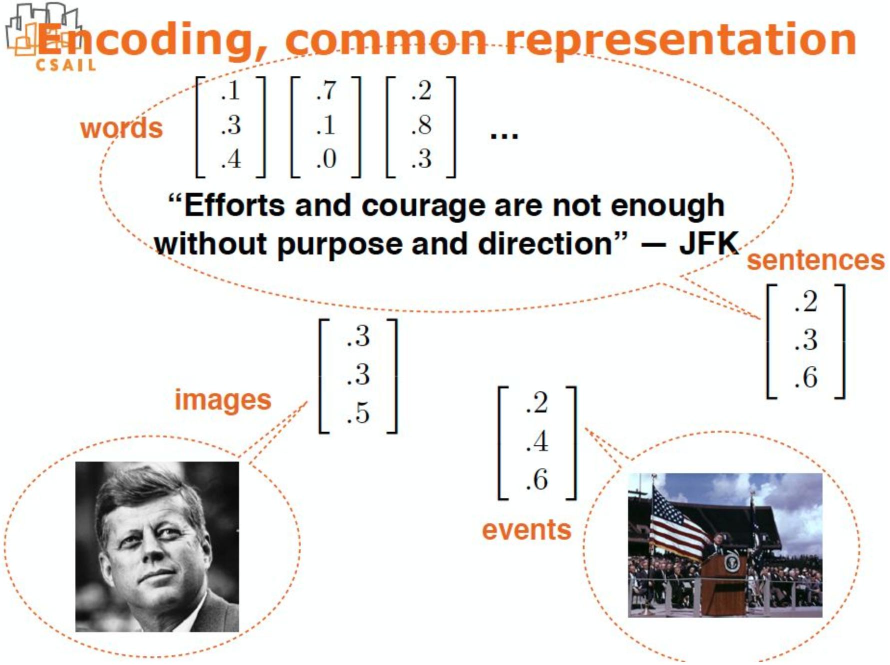
The field of natural language processing (NLP) has been transformed by massive pre-trained language models. They form the basis of all state-of-the-art systems across a wide range of tasks and have shown an impressive ability to generate fluent text and perform few-shot learning. At the same time, these models are hard to understand and give rise to new ethical and scalability challenges. In this course, students will learn the fundamentals about the modeling, theory, ethics, and systems aspects of large language models, as well as gain hands-on experience working with them.

https://stanford-cs324.github.io/winter2022/

https://microsoft.github.io/generative-ai-for-beginners/#/

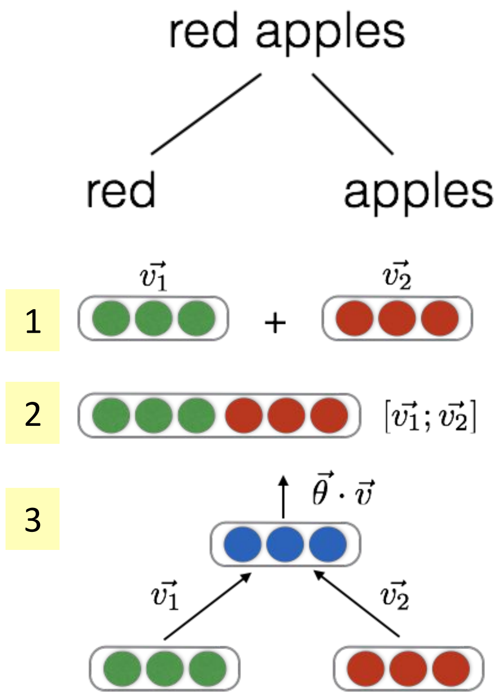
Reasons for exploring Deep Learning

- Learned features are easy to adapt and fast to learn
- Deep learning provides a very flexible, universal, and learnable framework for representing world, visual, and linguistic information

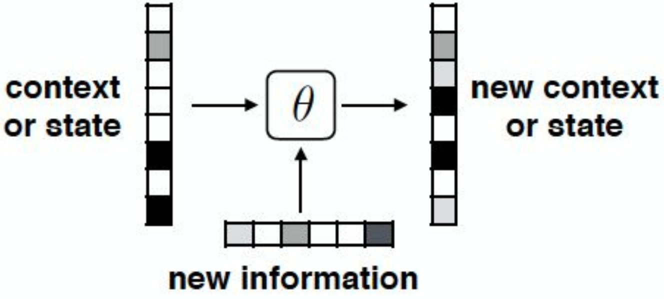


Reasons for exploring Deep Learning (cont.)

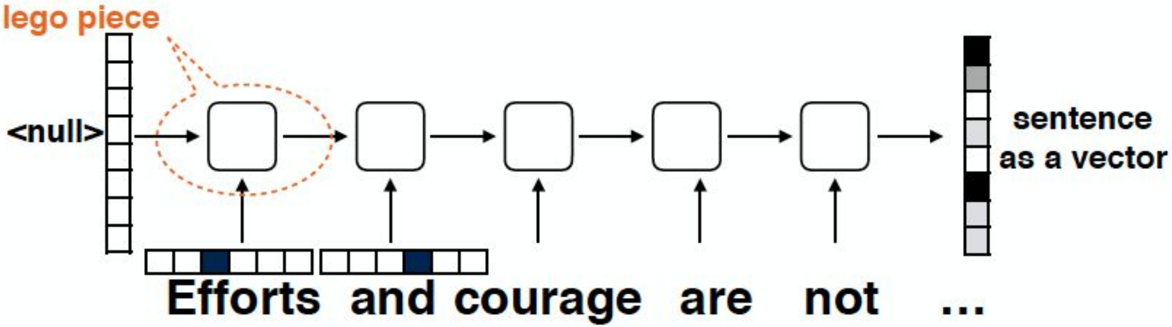
- Flexible neural "Lego pieces"
  - Common representation, diversity of architectural choices
- Can represent any levels of NLP
  - Word
  - Phrase
  - Sentence
  - Paragraph (document)



Reasons for exploring Deep Learning (cont.)
Example of encoding sentences



RNN
(LSTM, GRU)



Current problems with DL-based solutions

- Not so transparent
    - Bias unintentionaly learned from data
    - Blackbox: Don’t know when it will fail, how it will fail, hard to fix if it’s wrong

หมูกรอบ => |หมู|กรอบ|
ข้าวผัดคะน้าหมูกรอบหนึ่งจาน => |ข้าวผัด|คะน้า|หมูก|รอบ|หนึ่ง|จาน|

- Resource intensive (data and compute)

3) NLP System Building Overview


Advanced NLP Spring 2024

Natural language processing technology attempts to model human language with computers, tackling a wide variety of problems from automatic translation to question answering. CS11-711 Advanced Natural Language Processing (at Carnegie Mellon University's Language Technology Institute) is an introductory graduate-level course on natural language processing aimed at students who are interested in doing cutting-edge research in the field. In it, we describe fundamental tasks in natural language processing such as syntactic, semantic, and discourse analysis, as well as methods to solve these tasks. The course focuses on modern methods using neural networks, and covers the basic modeling and learning algorithms required therefore. The class culminates in a project in which students attempt to reimplement and improve upon a research paper in a topic of their choosing.

Carnegie Mellon University

Language Technologies Institute

https://phontron.com/class/anlp2024/

A General Framework for NLP Systems

Create a function to map an input *X* into an output *Y*, where *X* and/or *Y* involve language

Input X	Output Y	Task
Text	Continuing Text	Language Modeling
Text	Text in Other Language	Translation
Text	Label	Text Classification
Text	Linguistic Structure	Language Analysis
Image	Text	Image Captioning

Methods for Creating NLP Systems

- **Rules**: Manual creation of rules

```
def classify(x: str) -> str:
    sports_keywords = ["baseball", "soccer", "football", "tennis"]
    if any(keyword in x for keyword in sports_keywords):
        return "sports"
    else:
        return "other"
```

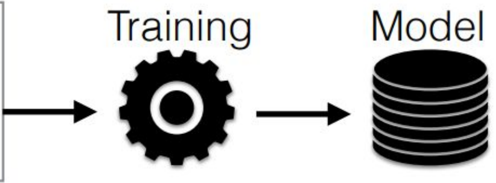
- **Prompting**: Prompting a language model w/o training

> If the following sentences is about "sports" reply "sports". Otherwise reply "other".
>
> {X}

→ LM

- **Fine-tuning**: Machine learning from paired data <X, Y>

I love to play baseball.	sports
The stock price is going up.	other
He got a hat-trick yesterday.	sports
He is wearing tennis shoes.	other



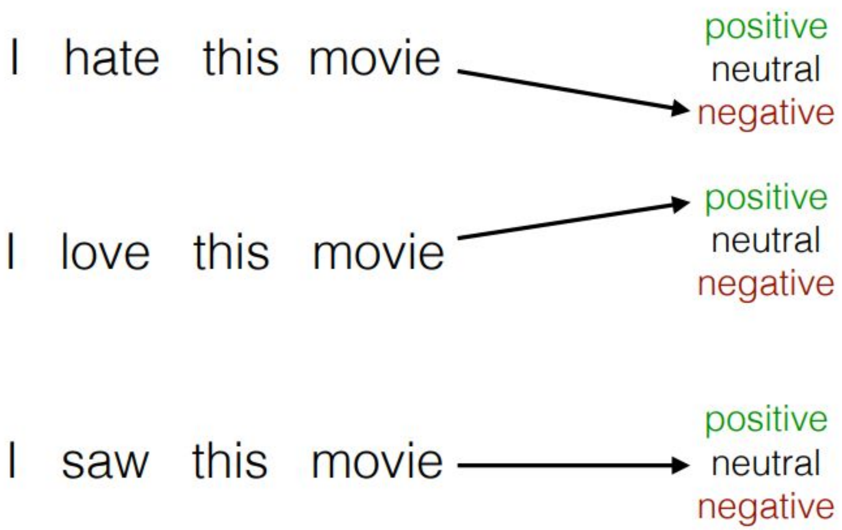
From https://phontron.com/class/anlp2024/

Data Requirements for System Building

- **Rules/prompting based on intuition**: No data needed, but also no performance guarantees
- **Rules/prompting based on spot-checks:** A small amount of data (small testing data) with input X only
- **Rules/prompting with rigorous evaluation**: Development set with input X and output Y (e.g., 200-2000 examples). Additional held-out test sets are also preferable (full testing data).
- **Fine-tuning**: Additional train set. More is often better — constant accuracy increases when data size doubles (training & testing data).

Example Task: Review Sentiment Analysis

Given a review on a reviewing website (*X*), decide whether its label (*Y*) is positive (1), negative (-1) or neutral (0)



From https://phontron.com/class/anlp2024/

A Three-step Process for Making Predictions

- **1) Feature extraction**: Extract the salient features for making the decision from text
- **2) Score calculation (model)**: Calculate a score for one or more possibilities
- **3) Decision function**: Choose one of the several possibilities

Formally

1 **Feature Extraction:** $\mathbf{h} = f(\mathbf{x})$

2 **Score Calculation:** binary, multi-class

$$s = \mathbf{w} \cdot \mathbf{h} \qquad \mathbf{s} = W\mathbf{h}$$

3 **Decision:** $\hat{y} = \text{decide}(\mathbf{s})$

From https://phontron.com/class/anlp2024/

Demo: Sentiment Classification

- **Rule-Based Model**
- ML-Based Model (BOW)
- NN-Based Model (not now)

Demo1: Rule-based Sentiment Classification Code Walk!

See code for all major steps:

1. Featurization
2. Scoring
3. Decision rule
4. Accuracy calculation
5. Error analysis

From https://phontron.com/class/anlp2024/

Now let's improve!

1. What's going wrong with my system? $\longrightarrow$ Look at error analysis
2. Modify the system (featurization, scoring function, etc.)
3. Measure accuracy improvements, accept/reject change
4. Repeat from 1
5. Finally, when satisfied with dev accuracy, evaluate on test!

Rule-based Challenges

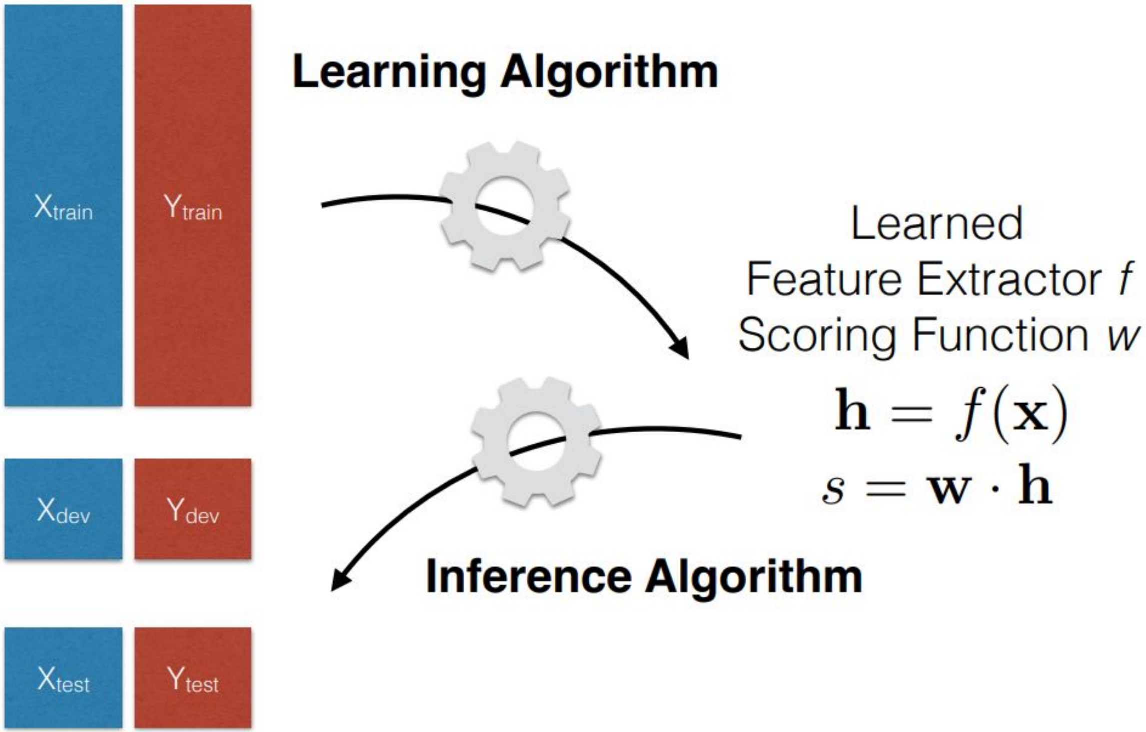
Challenge	Examples	Solution
Low Frequency Words	The action switches between past and present, but the material link is too **tenuous** to anchor the emotional connections that **purport** to span a 125-year divide . → negative	Keep working till we get all of them? Incorporate external resources such as sentiment dictionaries?
Conjugation	It's basically an **overlong** episode of Tales from the Crypt . → negative	Use the root form and POS of the word? *Note: Would require morphological analysis.*
Negation	This one is **not** nearly as **dreadful** as expected . → positive	If a negation modifies a word, disregard it. Note: Would probably need to do syntactic analysis.
Metaphor, Analogy	Puts a human face on a land most Westerners are unfamiliar with. → positive	???

From https://phontron.com/class/anlp2024/

Demo: Sentiment Classification

- Rule-Based Model
- **ML-Based Model (BOW)**
- NN-Based Model (not now)

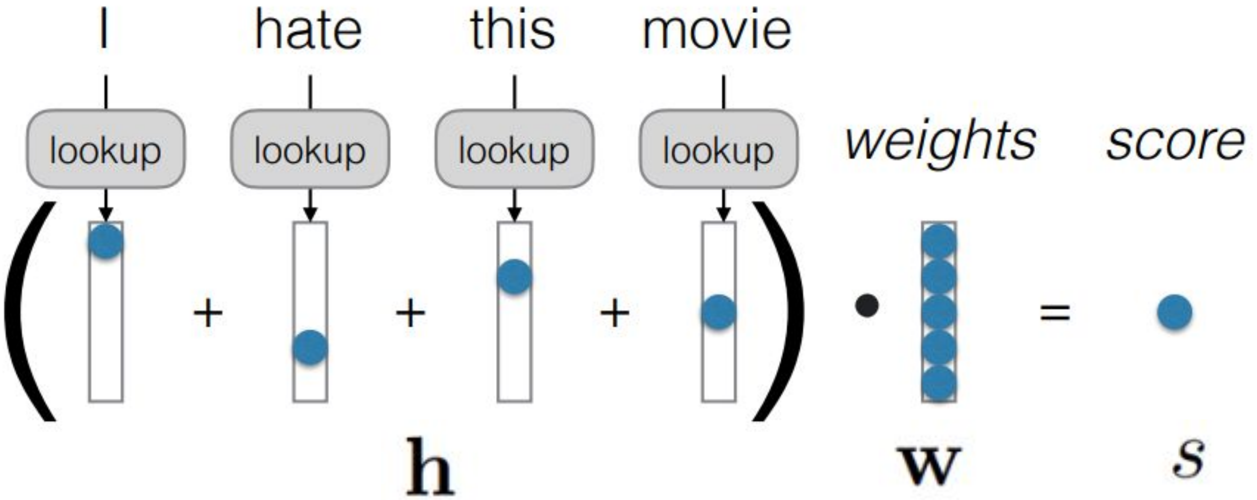
Machine Learning-Based NLP



From https://phontron.com/class/anlp2024/

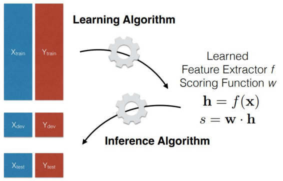
A First Attempt: Bag of Words (BOW)

Aim to solve low freq. words



Extracted handcrafted features
(OneHot; Dict)

Learnable weight
(ML)



Features *f* are based on word identity, weights *w* learned

Which problems mentioned before would this solve?

From https://phontron.com/class/anlp2024/

What do Our Vectors Represent?

- **Binary classification**: Each word has a single scalar, positive indicating “yes” and negative indicating "no.”

love	2.4
hate	-3.5
nice	1.2
no	-0.2
dog	-0.3
…	…

From https://phontron.com/class/anlp2024/

Demo2: Simple Training of BOW Models

Using an algorithm called “structured perceptron”

```
feature_weights = {}
for x, y in data:
    # Make a prediction
    features = extract_features(x)
    predicted_y = run_classifier(features)
    # Update the weights if the prediction is wrong
    if predicted_y != y:
        for feature in features:
            feature_weights[feature] = (
                feature_weights.get(feature, 0) +
                y * features[feature]
            )
```


From https://phontron.com/class/anlp2024/

What's Still Missing in BOW?

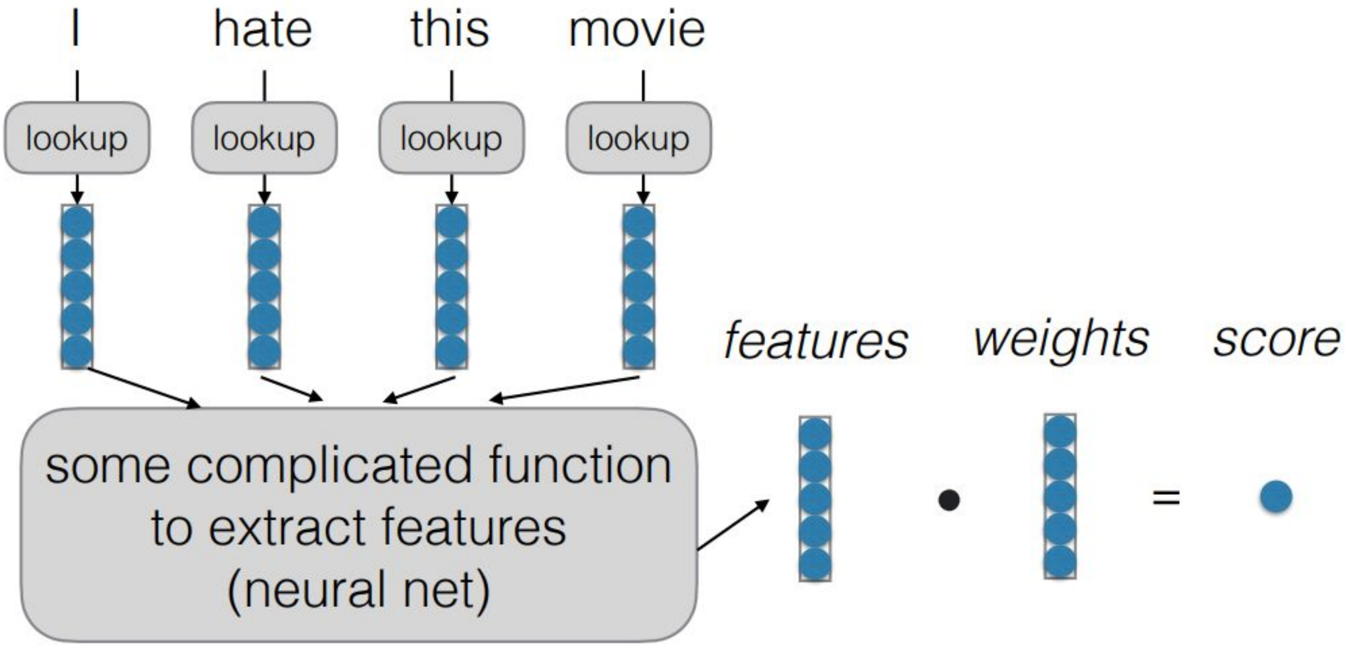
- Handling of *conjugated or compound words*
  - I **love** this move -> I **loved** this movie
- Handling of *word similarity*
  - I **love** this move -> I **adore** this movie
- Handling of *combination features*
  - I **love** this movie -> I **don't love** this movie
  - I **hate** this movie -> I **don't hate** this movie
- Handling of *sentence structure*
  - It has an interesting story, **but** is boring overall

From https://phontron.com/class/anlp2024/

Demo: Sentiment Classification

- Rule-Based Model
- ML-Based Model (BOW)
- **NN-Based Model (not now)**

A Better Attempt: Neural Network Models?



Powerful enough to perform classification, LM, any task!

From https://phontron.com/class/anlp2024/

4) This Course

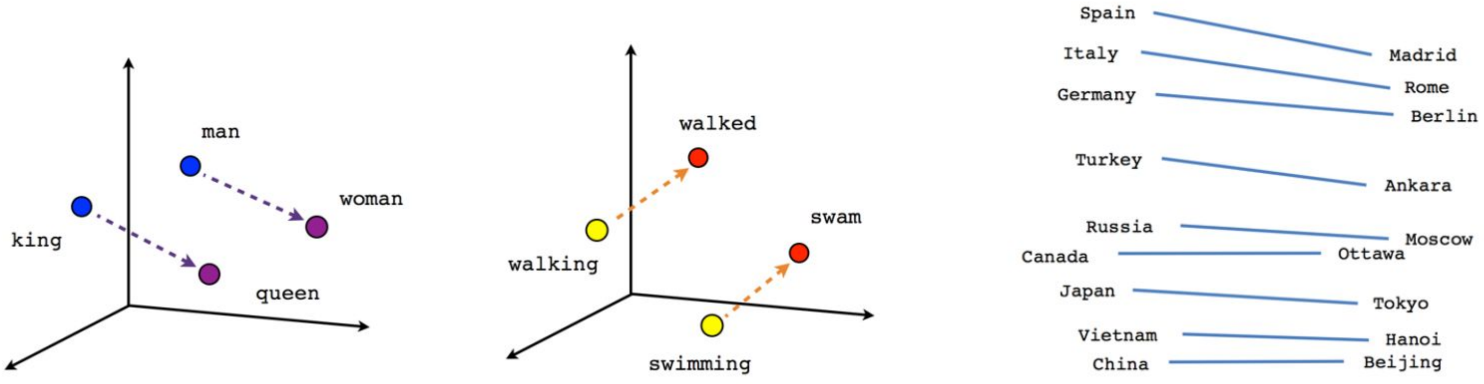
- Thai NLP Challenges
- Class Schedule
- Class Grading

In this class, we ask:

- What goes into the building blocks of **state-of-the-art NLP systems** that **work well** at some tasks?
- Where and why do current state-of-the-art NLP systems still **fail**?
- How can we **utilize NLP in real-world applications** given real-world challenges?

Some NLP Challenges

- Complexity in representing, learning and using linguistic/situational/world/visual knowledge



Male-Female

Verb tense

Country-Capital

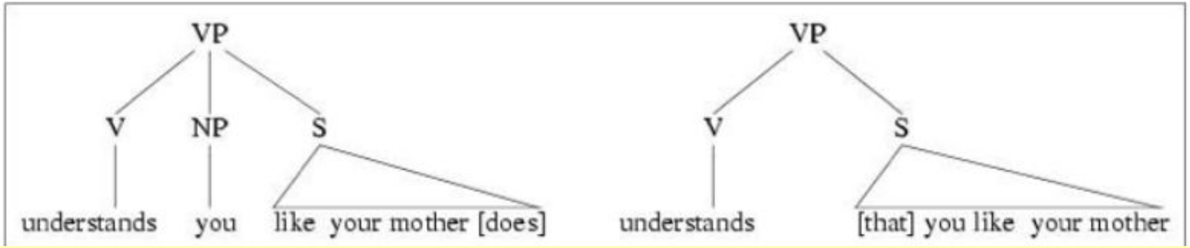
Ref: Prof. Christopher Manning, CS224N/Ling284, 2017

Some NLP Challenges (cont.)

- Human languages are ambiguous (unlike programming and other formal languages), so some parts can be ignored.
- Human languages are interpretation that depend on real world, common sense, and contextual knowledge (pragmatic analysis)

At last, a computer understands you like your mother."

Ambiguity at syntactic level: Different structures lead to different interpretations



The Pope's baby steps on gays. [Ref: Prof. Christopher Manning, CS224N/Ling284, 2017]

Ref: Prof. Christopher Manning, CS224N/Ling284, 2017 Ref: Prof. Regina Barzilay, NLP @MIT

Thai NLP Challenges

Word segmentation

- No word delimiters
- ฉัน|นำ|ดอก|ไม้|ไป|ไหว้|ศาล|พระ|ภูมิ|ที่|โรง|เรียน|ประจำ|
- ฉัน|นำ|ดอกไม้|ไป|ไหว้|ศาลพระภูมิ|ที่|โรงเรียน|ประจำ|
- ฉัน|นำ|ดอก|ไม้|ไป|ไหว้|ศาล|พระภูมิ|ที่|โรง|เรียน|ประจำ|
- ฉัน|นำ|ดอกไม้|ไป|ไหว้|ศาลพระภูมิ|ที่|โรงเรียนประจำ|

Sentence segmentation

- No sentence boundary markers

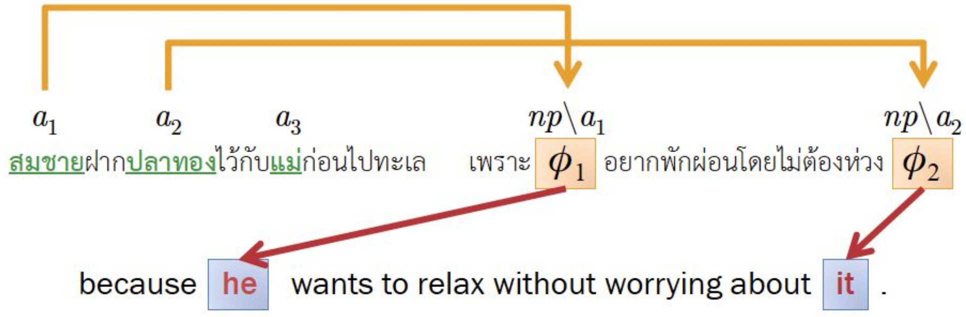
อย่างไรก็ตาม อดีตประธาน ทปอ. กล่าวว่า มีการทักท้วงเรื่องนี้มาตลอดว่า มี
ช่วงเวลาว่างนานขนาดนี้ ทำไมถึงยังต้องมีการจัดสอบนอกเหนือจากนี้อีก
เพราะการสอบล่วงล้ำไปในเวลาระหว่างเรียนมัธยมนั้นกระทบกับเรื่องอื่นๆ
โดยเฉพาะการเรียนในชั้นเป็นวงจรลูกโซ่ แนวโน้มที่เข้ามาแก้เรื่องนี้ เป็น
ความคิดที่ดี แต่ยังไม่เห็นเรื่องใช้ผลการเรียนในชั้นมาเป็นองค์ประกอบรับตรง
ซึ่งอาจทำให้เด็กไม่สนใจห้องเรียน และมุ่งกวดวิชา ทำให้การสอบเข้าอุดมศึกษา
ตกเป็นจำเลยข้อหาทำลายระบบการศึกษาขั้นพื้นฐาน วนไปสู่ปัญหาเก่าๆ ได้

Ref: Introduction to Thai NLP (Prachya Boonkwan), 2017

Thai NLP Challenges (cont.)

Syntax ambiguity

- Pronouns and some constituents can be omitted as long as they can be implied from the context



Nostalgic Thai slang

Ref: Introduction to Thai NLP (Prachya Boonkwan), 2017

อัปเดต 60 ศัพท์วัยรุ่น 2567 ตัวตึง รู้ไว้ไม่งงเขาคุยอะไรกัน

ผู้เขียน : Sale Here Editor

12 ก.ย. 67 (วันที่อัพเดตล่าสุด)

https://salehere.co.th/articles/update-teen-words

คำศัพท์	คำแปล
ตัวตึง	ตัวท็อป , ตัวเต็ง
ตัวแม่	มีความมั่นใจ เชิ่ด เก๋
บูด	ไม่เริ่ด ,ไม่ถูกใจ ,ไม่ดี
ความโป๊ะเป็นศูนย์	ไม่มีที่ติ
ตุย	เหนื่อย ,ไม่แรง
จ้อจี้	โกหก
เฟื่อน	สถานะไม่ชัดเจนระหว่างเพื่อนกับแฟน
ว่าซ่าน	อ๋อว่างั้น ,ประมาณนั้น
จะแล้วไหม	จะเสร็จไหม ,เสร็จหรือยัง
ฟาด	จัดการทันที ,สวย
ตำ	ไปซื้อ
เอาหมดไม่สนว่าลูกใคร	เอาหมดไม่ว่าของสิ่งนั้นเป็นของใคร
สู่ขิต	เพี้ยนมาจาก สู่สุคติ
จึ้ง	ดีมาก ,สวย ,เริ่ด
มาเอาไร	มาทำอะไร ,มาเอาอะไร
อย่าหาทำ	มาทำอะไร ,มาเอาอะไร
หยุดดิ (เสียงสูง)	อย่าทำ ,หยุด ไม่ให้ทำ
แบบตะโกน	ดีที่สุด ,ดีมากกว่าปกติจนต้องตะโกน

Class Schedule ttps://github.com/ekapolc/NLP_2025

#	Description
1	Intro; Tokenization (traditional+deep)
2	Language Model (N-grams, Smoothing, Perplexity, CE loss)
3	Attention + transformer + architectures (BERT, GPT, etc.)
4	Token classification
5	sentence representation (traditional+deep/contrastive learning)
6	sequence classification
7	Text generation (beamsearch, top-k top-p, metrics) + BPE subword + evaluation
8	LLM take home exam (kaggle) + Project Announcement + Paper Announcement
	Midterm Exam Week (3-7 Mar 2025) - no class
9	Midterm Exam (paper exam - in class)
10	Prompting/Lora/Adaptation techniques
11	Recent topics (bias, benchmark, RAG, agentic, test-time compute)
12	Paper Presentation & Progress Report
	No class
13	Deployment
14	Multimodal
	Final Exam Week (28 Apr -14 May 2025); No Final Exam
	Project Presentation

Course Grading

- Assignments 35% (4%x10 HW capped at 35%)
- Midterm 35% (in class exam 25% + take home 10%)
- Paper presentation 10%
- Project 20%

Speech and Language Processing (3rd ed. draft)

Dan Jurafsky and James H. Martin

https://web.stanford.edu/~jurafsky/slp3/

3rd edition draft as of August 20, 2024 release

5) NLP Tools

Implementation

python™

NumPy

scikits learn
machine learning in Python

K + TensorFlow

PyTorch

PyTorch Lightning

Google colab

NLP Libraries

NLTK 3.5 documentation

NEXT | MODULES | INDEX

nltk 3.9.1

```
pip install nltk
```

Natural Language Toolkit

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.

spaCy   USAGE   MODELS   API   UNIVERSE   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,

Blazing fast

spaCy excels at large-scale information extraction tasks. It's written from the ground up in carefully memory-managed

Deep learning

spaCy is the best way to prepare text for deep learning. It interoperates seamlessly with TensorFlow, PyTorch, scikit-learn,

HUGGING FACE

On a mission to solve NLP, one commit at a time.

Star 39,335

Search model   t: Most downloads

- text-classification 📖
- token-classification 🏷️
- question-answering ❓
- multiple-choice 🔘
- masked-lm 🔥
- causal-lm 🍫
- summarization 📋
- translation 🇺🇳
- conversational 💬

Version updated as of 2 Jan 2025

https://pythainlp.org/docs/5.0/

Thai NLP Libraries

PyThaiNLP
dev (e034c54)
Published date: 16/05/2024

Search docs

NOTES

- FAQ
- Command Line
- Getting Started
- Tutorial Notebooks
- Installation
- FAQ
- License

PACKAGE REFERENCE:

- pythainlp.ancient
- pythainlp.augment
- pythainlp.benchmarks
- pythainlp.chat
- pythainlp.classify
- pythainlp.coref
- pythainlp.corpus
- pythainlp.el
- pythainlp.generate
- pythainlp.khavee
- pythainlp.morpheme
- pythainlp.parse
- pythainlp.phayathaibert
- pythainlp.soundex
- pythainlp.spell

/ PyThaiNLP documentation

PyThaiNLP documentation

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

Website: PyThaiNLP.github.io

Notes

- FAQ
- Command Line
- Getting Started
- Tutorial Notebooks
- Installation
- FAQ
- License

Package reference:

- pythainlp.ancient
- pythainlp.augment
- pythainlp.benchmarks
- pythainlp.chat
- pythainlp.classify
- pythainlp.coref
- pythainlp.corpus
- pythainlp.el
- pythainlp.generate
- pythainlp.khavee

Project description

PyThaiNLP is a Python library for Thai natural language processing. The library provides functions like word tokenization, part-of-speech tagging, transliteration, soundex generation, and spell checking.

Version	Description	Status
5.0.5	Stable	Change Log
dev	Release Candidate for 5.1	Change Log

Version updated as of 2 Jan 2025

LLM tools

HUGGING FACE

LlamaIndex

LangChain

LangSmith

LangGraph

Any questions? :)