

Introduction to Natural Language Processing

2110572: NLP SYS

Assoc. Prof. Peerapon Vateekul, Ph.D.

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?

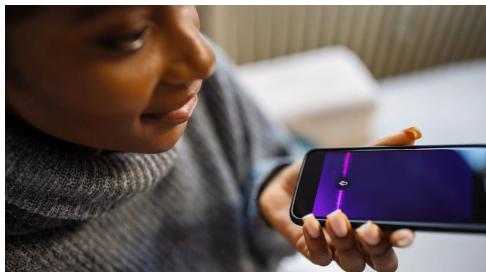
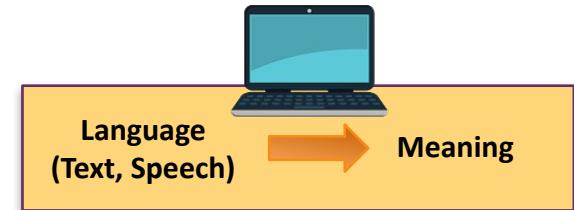
- Definition & Levels of Understanding
- NLP today

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)



NLP (Interpretation)

I want to eat Japanese food.

May I order Yayoi for you?

AI/ML (Generation)

What can I help you with?



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

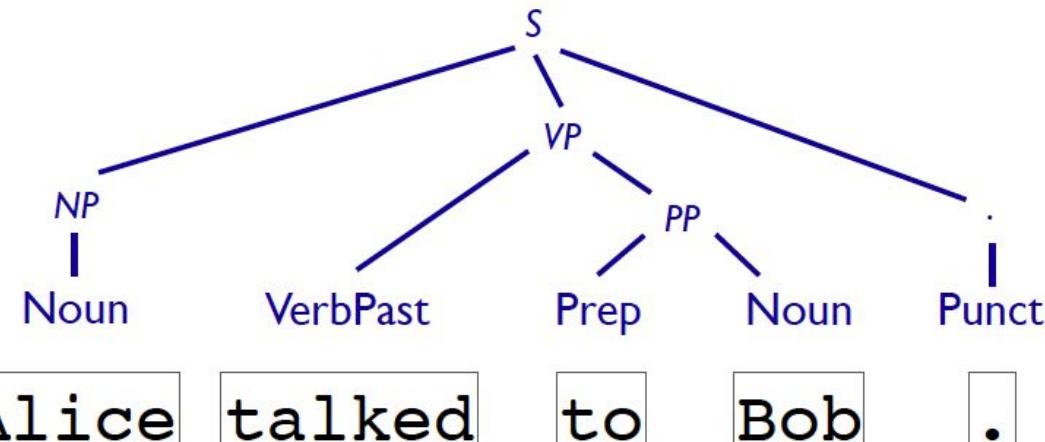
CommunicationEvent(e)

Agent(e, Alice)

Recipient(e, Bob)

SpeakerContext(s)

TemporalBefore(e, s)



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

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

Global Domination
Home Museum Demolished Kabul Airport Expanded Regional Role
Mandela Served Sentences Nelson Mandela's Japan European Union
Central African Republic
South Africa Lebanon China Yangtze River
Nelson Mandela United States
Amritsar Supreme Court Nelson Mandela's
Britain Nelson Mandela's
David Cameron North Syria Shanghai Syrians Punjab
Photographer Says Syria Pretoria Defense Panel Calls

Application

Tokenization

- Input: ขสมก. เลี้ง จัดหารถ
- Output: ขสมก., เลี้ง, จัดหารถ

Part of Speech tagging

- Input: [ขสมก., เลี้ง, จัดหารถ]
- Output: [(ขสมก., NR), (เลี้ง, VV), (จัดหาร, VV), (รถ, NN)]

NER

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

Application

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

PENN Part Of Speech Tags

- NR – proper noun
 - VV - Main verbs in clauses, verb-form
 - NN – Non-proper noun
- Ref: BEST2010 dataset

Named Entity Tags

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



1.2) NLP Applications

NLP can Answer Questions

1

Who is Peerapon Vateekul?



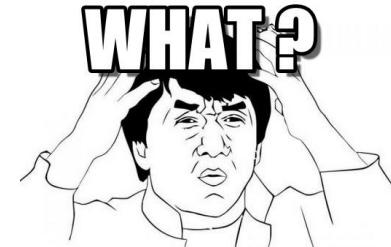
Asst. Prof. Dr. Peerapon Vateekul is a faculty member in the Department of Computer Engineering at Chulalongkorn University in Bangkok, Thailand. CHULALONGKORN UNIVERSITY He earned his Ph.D. in Electrical and Computer Engineering from the University of Miami in 2012. CHULALONGKORN UNIVERSITY His research 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



...but (sometimes) makes up facts



©meganfr

2

What's special about Peerapon Vateekul'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.

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

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:

Warongpon Phuenpathorn
Chulalongkorn University

Pattamon Panyakaew
Chulalongkorn University

Peerapon Vateekul
Chulalongkorn University

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.

Here are three big themes to watch out for over the next 12 months.

Can the stock market rally go on? The new year has begun with stock markets in the UK and US hitting new record highs.

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.

เป็นปีใหม่ที่กำลังได้รับการแสดงความคิดเห็นของผู้เชี่ยวชาญให้มุมมองของพวากษาเกี่ยวกับสิ่งที่ 2018 เก็บไว้

ต่อไปนี้เป็นหัวข้อใหญ่สามข้อที่ควรระวังในช่วง 12 เดือนข้างหน้า

การซัมมูมตลาดหุ้นสามารถดำเนินต่อไปได้หรือไม่? ปีใหม่เริ่มมีตลาดหุ้นในสหรัฐอเมริกาจัดและสะท้อนพุ่งสูงเป็นประวัติการณ์

ตั้งนี้เฉลี่ยอุดสาหรูรวมดาวโจน斯ปรับตัวสูงขึ้นกว่า 25,000 จุดเป็นครั้งแรกในสัปดาห์นี้ ขณะที่ดัชนี S & P 500 ที่ใหญ่ขึ้นก็อยู่ในระดับสูงเป็นประวัติการณ์

Markets, Brexit and Bitcoin: 2018's themes

By Chris Johnston
Business reporter

5 January 2018

f t m 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) ▾

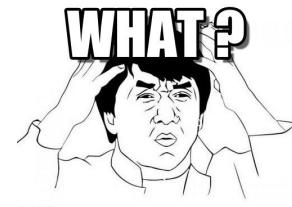
แล้วเธอจะเจอดี
Sarcasm

Then you will meet with good things.

ลำตัดคณะนี่เล่นถึงพริกถึงขิงจริง
Uncommon Words

Idioms

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



NLP asks “why” are we Searching

Google aquaman

All Images Videos News Maps More

About 164,000,000 results (0.69 seconds)

What is the query's intent?

Watch movie?

Showtimes for Aquaman
All times are in Thailand Time

Aquaman Movie Official Website - In theaters December 21, 2018
<https://www.aquamanmovie.com/>

Aquaman - #AquamanMovie- In theaters December 21st, 2018.

Aquaman (2018) - Rotten Tomatoes
https://www.rottentomatoes.com/m/aquaman_2018/

★★★★★ Rating: 64% - 298 reviews

Dec 21, 2018 - Critic Consensus: **Aquaman** swims with its entertainingly ludicrous tide, offering up CGI superhero spectacle that delivers energetic action with ...

Aquaman (film) - Wikipedia
[https://en.wikipedia.org/wiki/Aquaman_\(film\)](https://en.wikipedia.org/wiki/Aquaman_(film))

Aquaman is a 2018 American superhero film based on the DC Comics character of the same name, and distributed by Warner Bros. Pictures. It is the sixth ...

Amber Heard · James Wan · Ocean Master · Yahya Abdul-Mateen II

Webs related to query?

8:30pm

อควาแมน เจ้าสมุทร

พ.ศ. 2561 · ภาพยนตร์แนวแฟนตาซี/ภาพยนตร์รัตนวิทยา วันฉายาสดร. · 2 ชั่ว. 22 นาที

7.6/10 IMDb | 64% Rotten Tomatoes | 55% Metacritic

94% ชอบภาพยนตร์เรื่องนี้

จาก Google

อควาแมน เจ้าสมุทร เป็นภาพยนตร์รัฐบาลอเมริกันจากปี 2018 ที่กำกับโดย James Wan และเขียนบทโดย David Leslie Johnson-McGoldrick และ Will Beall และเรื่องราวโดย Jason Momoa ...

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	6/12/08	BIOBX-L	0	P		3	CALCS	N
32875	56	1	7/11/08	BIOBX-B	0	N		3		
2247	72	1	4/12/08	BIOBX-R	0	N		3		
45521	61	1	3/30/08	BIOBX-B	0	B		3	CALCS	S
48987	41	1	4/5/08	BIOBX-B	0	P		3	CALCS	N
4179	67	1	5/12/08	BIOBX-B	0	P		2	CALCS	N
24300	59	1	3/31/08	BIOBX-L	0	N		3		
67960	64	1	4/7/08	BIOBX-R	0	P		3	MASS	O
43283	61	W	7/21/08	BIOBX-B	0	B		3		
43319	51	1	4/7/08	BIOBX-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

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

NER by spaCy

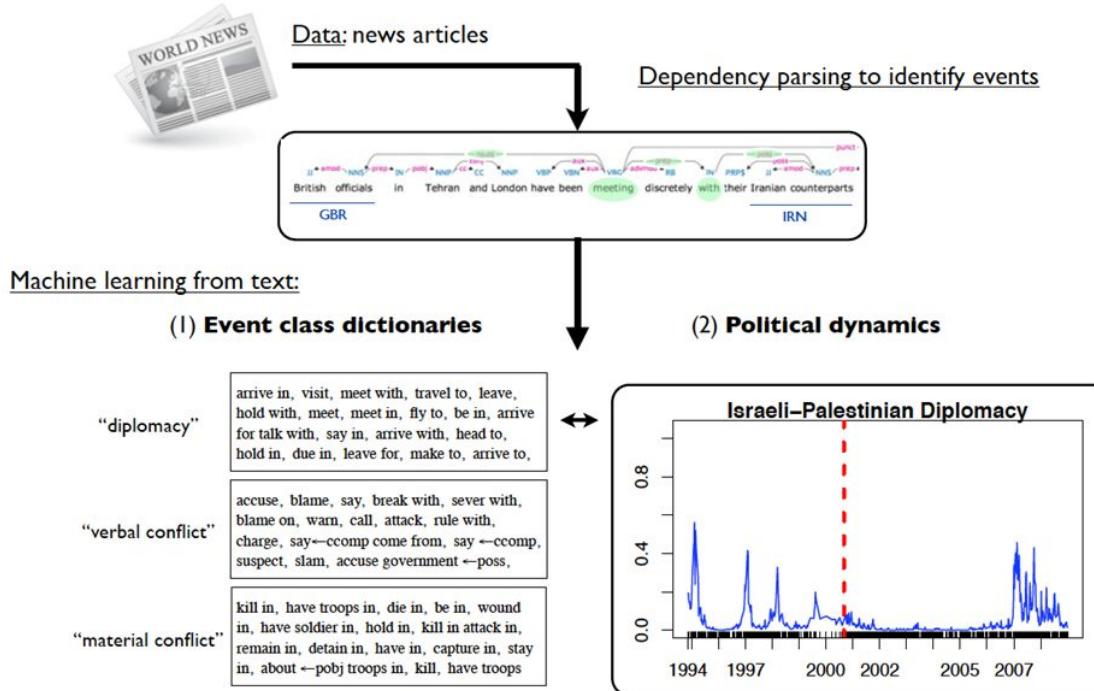
Suan Dusit PERSON Poll president Pornpan Buathong PERSON said the popularity of the Moo Deng PERSON story, which captivated much of the world, reflected the sentiment felt by Thais GPE this year DATE , made stressful by hard news like the Icon Group ORG case and the end of Move Forward ORG .

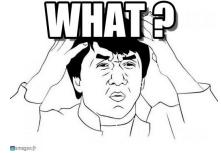
Different tools can be failed in various cases:

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



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."

Ref: Prof. Regina Barzilay, NLP @MIT



NLP is difficult!
Word-level ambiguity!

Many problems are trivialized in the LLM Era

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

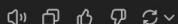


translate this to english pls "ແລ້ວເຮືອຈະເຈອດີ"

The translation of "ແລ້ວເຮືອຈະເຈອດີ" to English is:

"Then you'll get what you deserve" or "You'll face the consequences."

The tone can vary depending on the context, ranging from a lighthearted warning to a more serious or threatening implication.



will you ever get confused between the actress Hathaway and Berkshire Hathaway

It's unlikely! Anne Hathaway is a well-known actress, while Berkshire Hathaway is a multinational conglomerate led by Warren Buffett. The context usually makes it clear which one is being referred to. For example:

- If someone mentions movies or acting, it's Anne Hathaway.
- If it's about investments or business, it's Berkshire Hathaway.

If you're discussing both in the same conversation, I'll keep track of the context to avoid confusion.



2) History of NLP & Deep Learning

- History of NLP
- Deep Learning

A Brief Timeline of NLP

Logical Era

1950

Statistical Era

1980

Neural Era

1990

Neural (DL) Era

2010

- Alan Turing's Turing Test
- Chomsky's 'universal grammar'¹
- Interest in automatic translation²
- MIT ELIZA rule-based rephrasing
- MIT SHRDLU moving blocks
- Conceptual ontologies to chatterbots

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

- 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

- 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

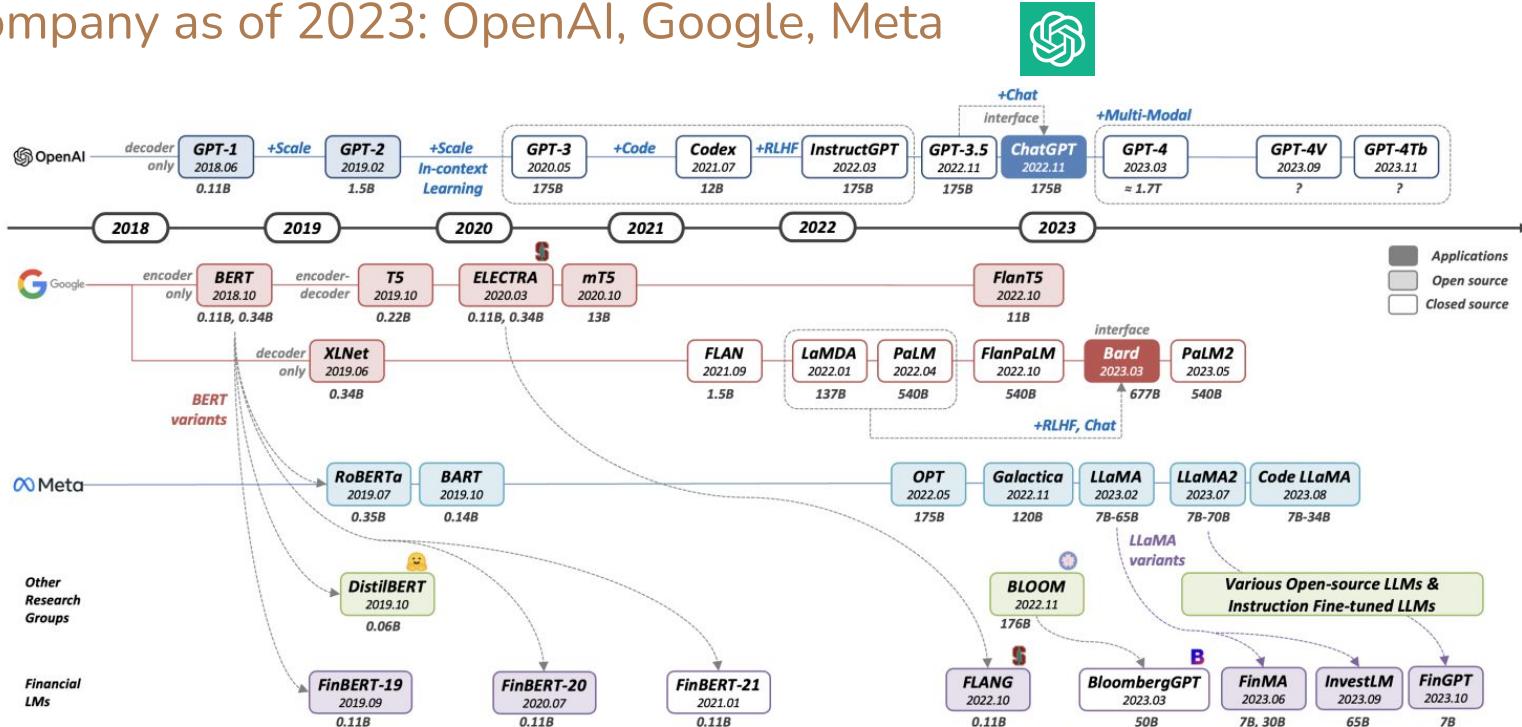


Figure 1: Timeline showing the evolution of selected PLM/LLM releases from the general domain to the financial domain.

A Brief Timeline of NLP since 2020 (LLM)

Based on timeline as of 2023 (overall companies)

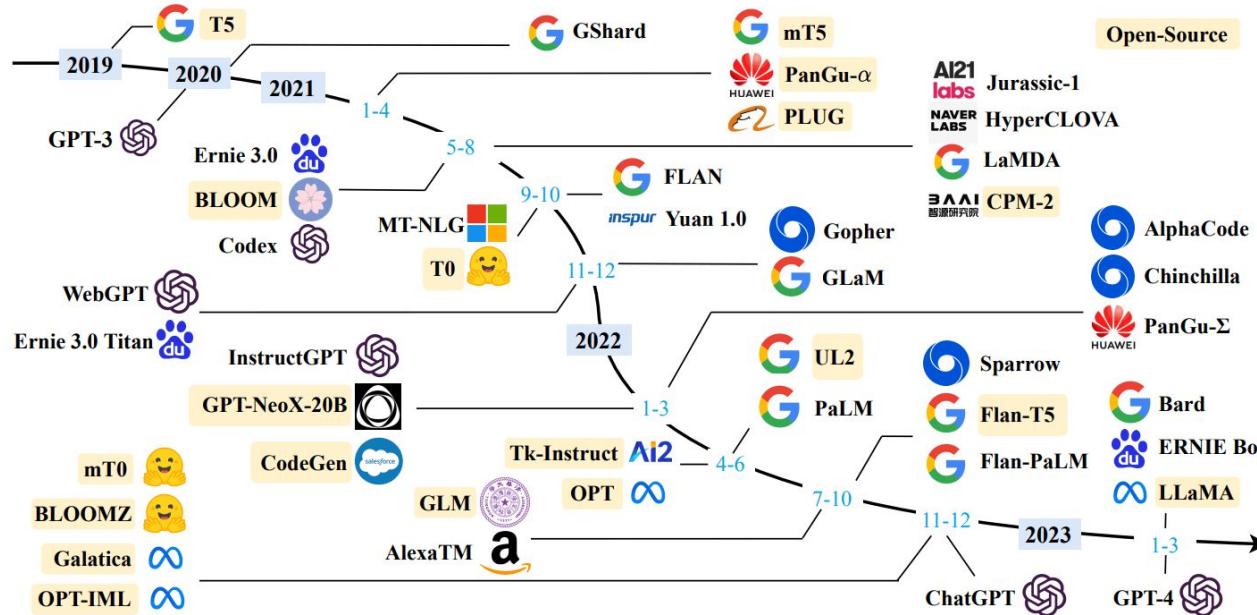


Fig. 1. A timeline of existing large language models (having a size larger than 10B) in recent years. We mark the open-source LLMs in yellow color.

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:

PICK UP A BIG RED BLOCK.
OK.

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?

A diagram showing several 3D blocks (cubes and pyramids) and two rectangular boxes. One box is labeled 'TOPCENTER' and the other 'TCENT'. Arrows indicate the movement of blocks between them.

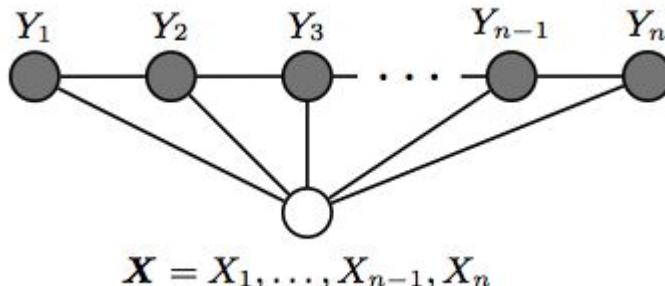
```
(DEFTHEOREM TC-GRASP
  (THCONSE (X Y)  (#GRASP $?X)
            (THGOAL(#MANIP $?X))
            (THCOND ((THGOAL (#GRASPING $?X)))
                     ((THGOAL (#GRASPING $_Y))
                      (THGOAL (#GET-RID-OF $?Y)
                             (THUSE TC-GET-RID-OF))))
            (T))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

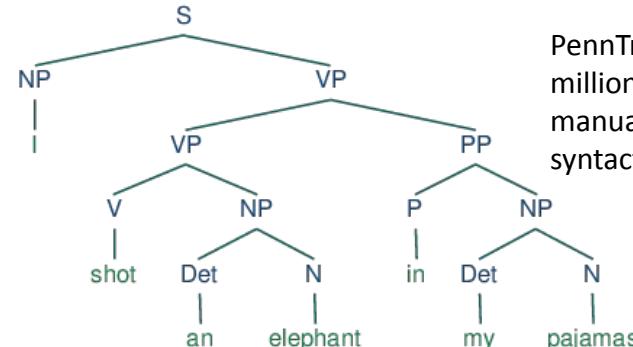
History: Statistical Era

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

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.intes2learn.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!**

Case Study: Determiner placement (cont.)

2) Statistical approach

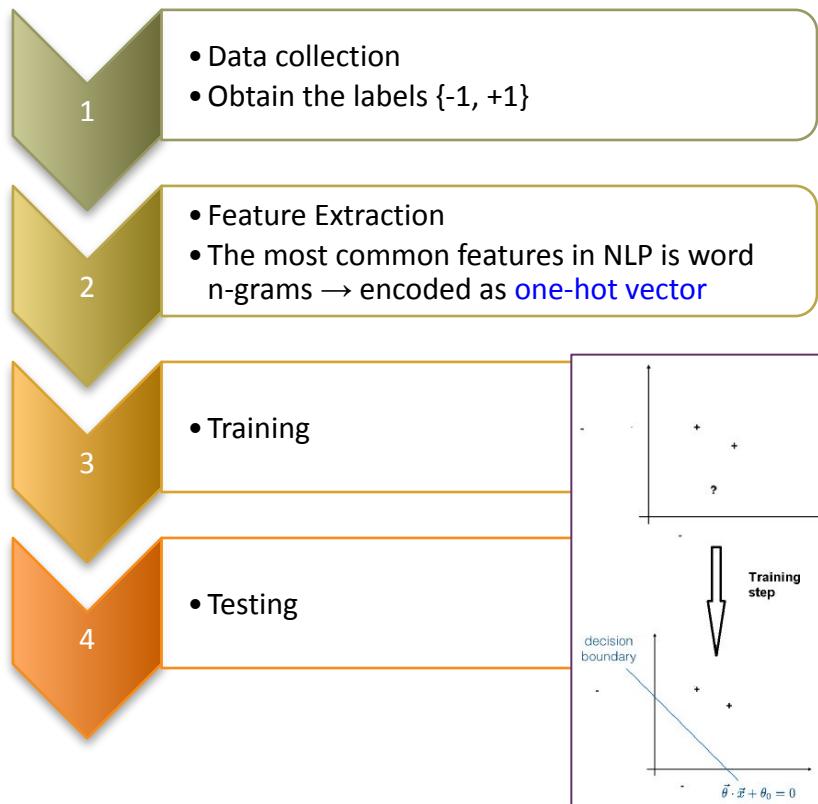
- Consider it as classification
- Predictions: {-1, +1}
- Features:
 - Plural?
 - first appearance in text?
 - head token
 - ...

“lazy monkeys”
[1 1 0 0 0 … 1]^T
↓
-1

“the United States”
[1 1 0 0 0 … 0]^T
↓
+1

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

Limitation of traditional statistical approach



- 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 → Deep Learning ***

pear

apple

$$\begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

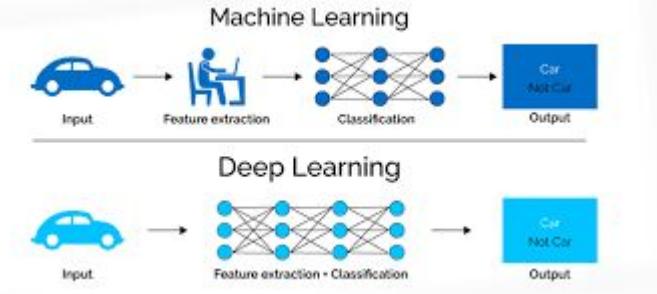
$$[0 \ 0 \ 1 \ 0 \ \dots \ 0]$$

↓

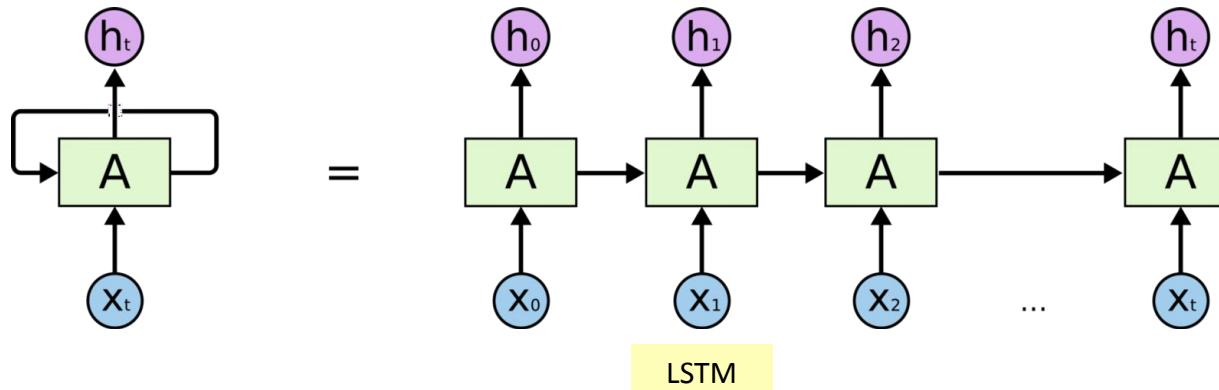
[0.6 0.2 0.3]

History: Neural Era

3) Deep Learning approach:



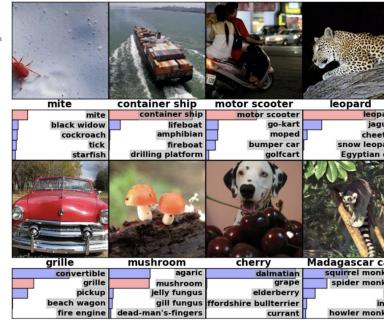
- It is a **feature-engineering embedded** neural approach.
- Since the 2010s, it has been gaining a lot of attention and showing many successes.



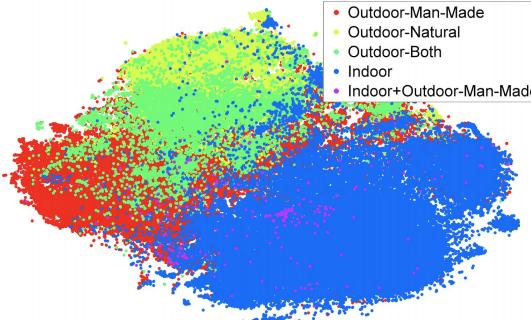
The Spark of DL in Various Domains: Success of AlexNet in Large Scale Visual Recognition Challenge (ILSVRC) 2012



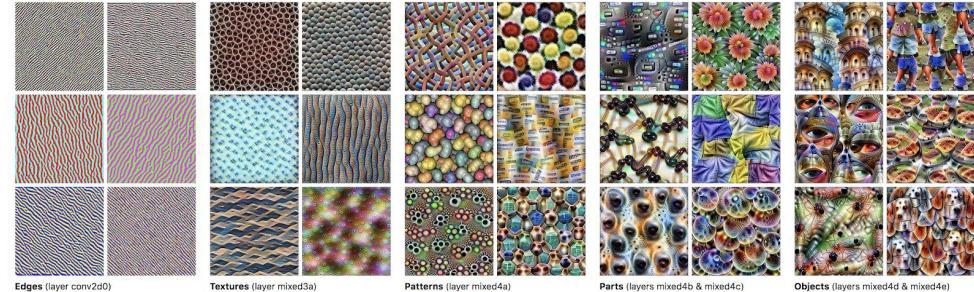
- 1,000 object classes (categories).
- Images:
 - 1.2 M train
 - 100k test.



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*.



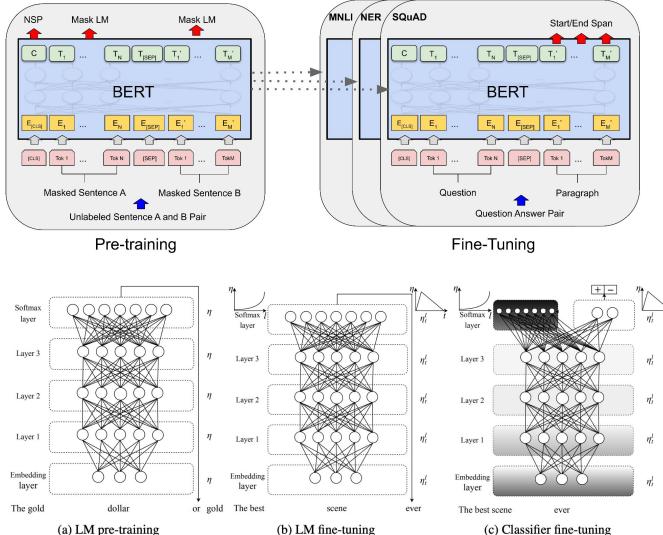
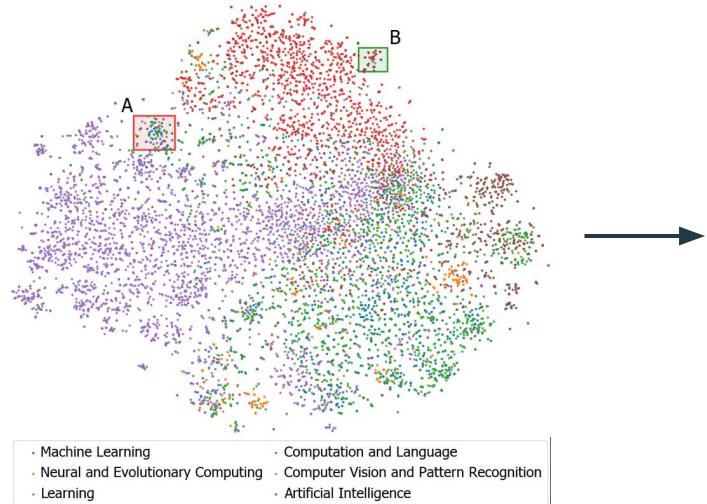
Features trained on ILSVRC-2012 generalize to the SUN-397 dataset. [Donahue et al., 2014]



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

NLP “Pretrained” Word Embeddings

Embeddings for arXiv papers (6 ML categories)



2013 Word2Vec GloVe

2018 Bert ULMFiT

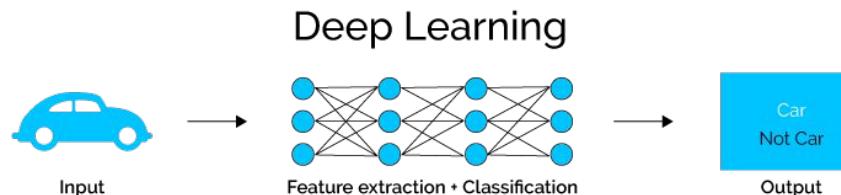
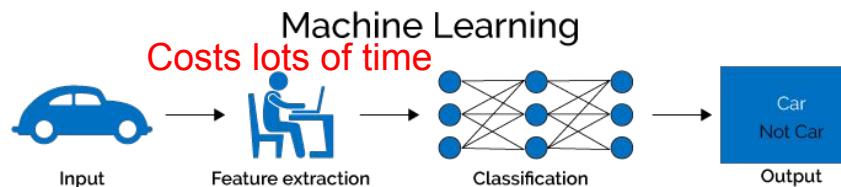
What is Deep Learning?



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



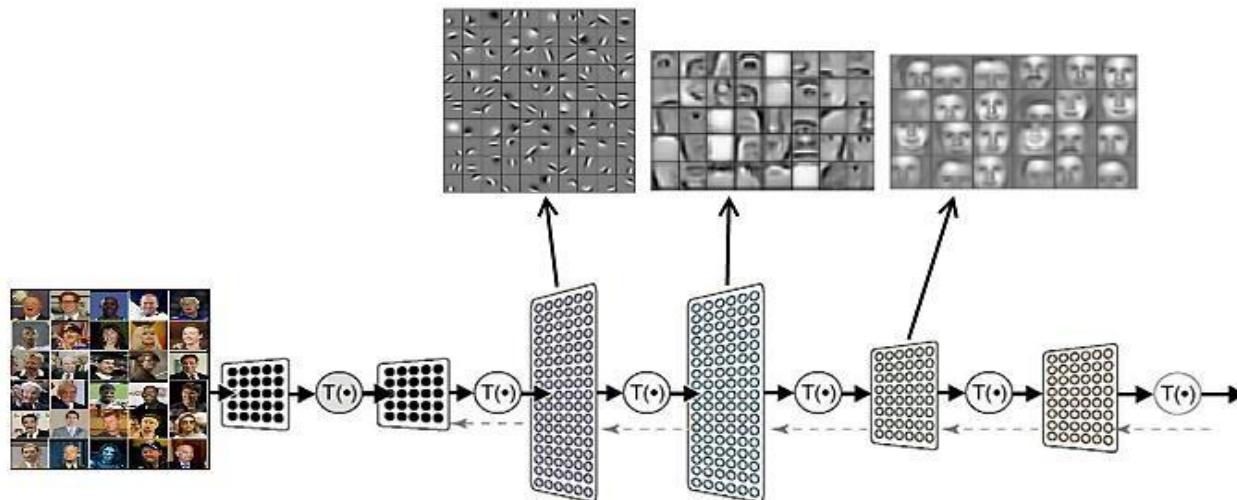
Utilizes learning algorithms that derive meaning out of data by using a hierarchy of multiple layers that mimic the neural networks of our brain.



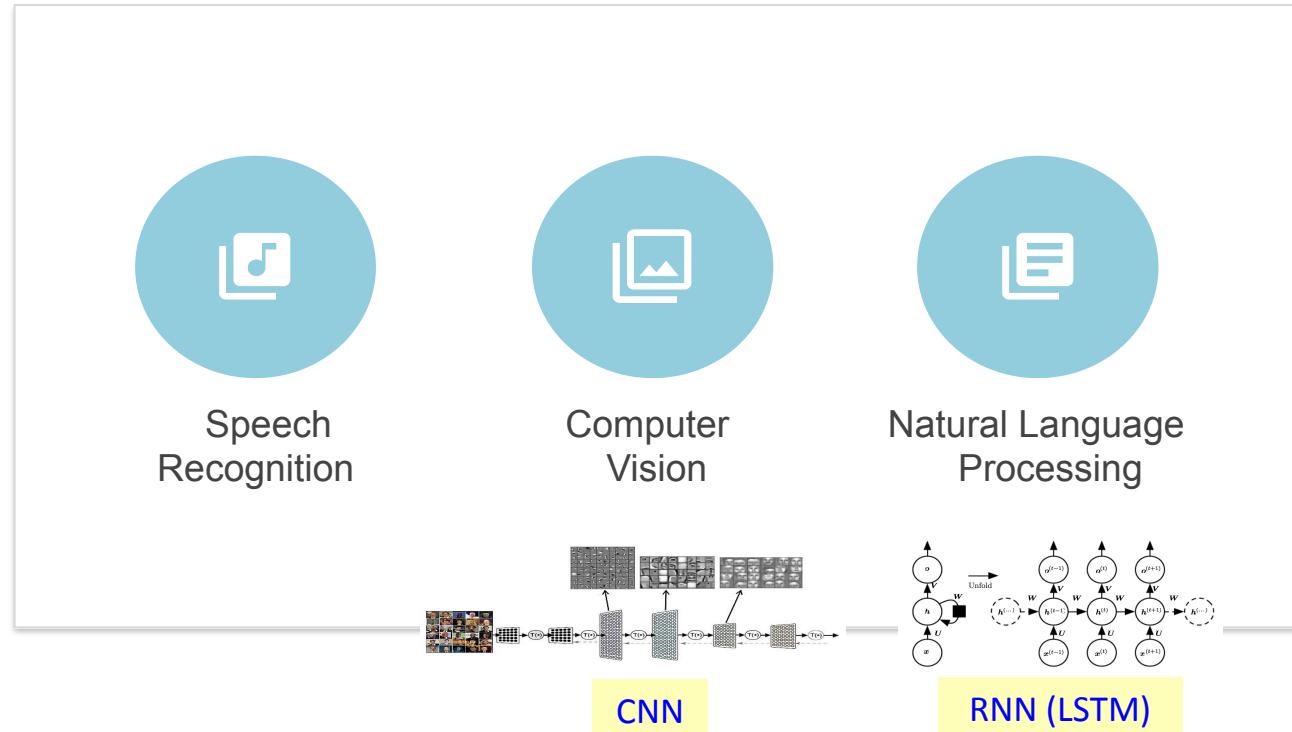
Deep Learning – Basics (cont.)

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



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.



CS224d: Deep Learning for Natural Language Processing



pucktada/cutkum

cutkum - Thai Word-Segmentation with Deep Learning in Tensorflow

วิชา NLP with Deep Learning ของ Stanford ของ Winter 2017 ล่าสุดครับ



Lecture Collection | Natural Language Processing with Deep Learning (Winter 2017) - YouTube

Natural language processing (NLP) deals with the key artificial intelligence technology of understanding...

YOUTUBE.COM

Thai word segmentation with bi-directional RNN

This is code for preprocessing data, training model and inferring word segment boundaries of Thai text with bi-directional recurrent neural network. The model provides precision of 99.04%, recall of 99.31% and F1 score of 99.18%. Please see the [blog post](#) for the detailed description of the model.

(Submitted on 16 Nov 2019)

AttaCut: A Fast and Accurate Neural Thai Word Segmente

Pattarawat Chormai, Ponrawee Prasertsom, Attapol Rutherford

Word segmentation is a fundamental pre-processing step for Thai Natural Language Processing. The current off-the-shelf solutions are not benchmarked consistently, so it is difficult to compare their trade-offs. We conducted a speed and accuracy comparison of the popular systems on three different domains and found that the state-of-the-art deep learning system is slow and moreover does not use sub-word structures to guide the model. Here, we propose a fast and accurate neural Thai Word Segmente that uses dilated CNN filters to capture the environment of each character and uses syllable embeddings as features. Our system runs at least 5.6x faster and outperforms the previous state-of-the-art system on some domains. In addition, we develop the first ML-based Thai orthographical syllable segmente, which yields syllable embeddings to be used as features by the word segmente.

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.

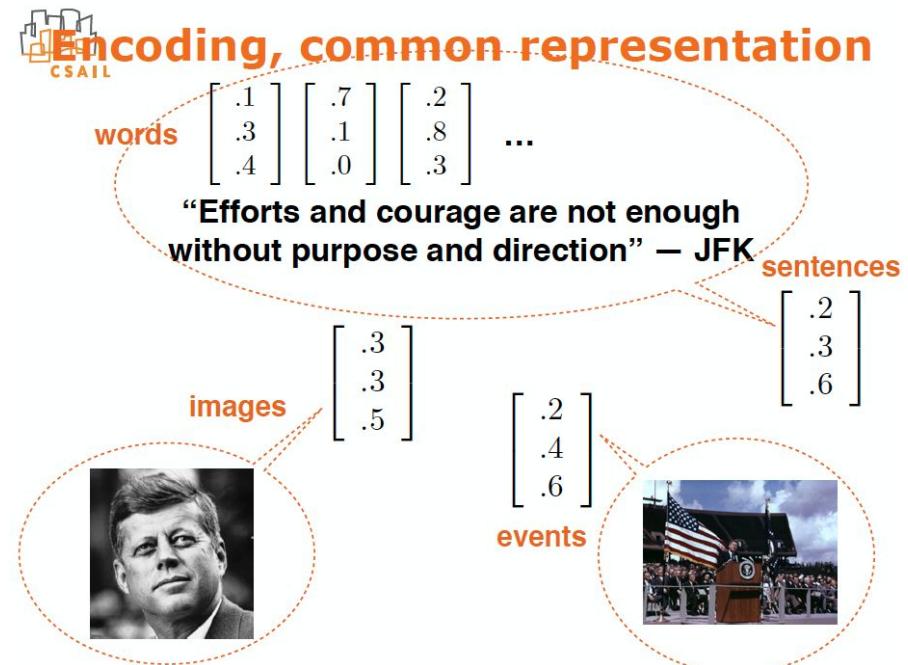
<https://stanford-cs324.github.io/winter2022/>



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

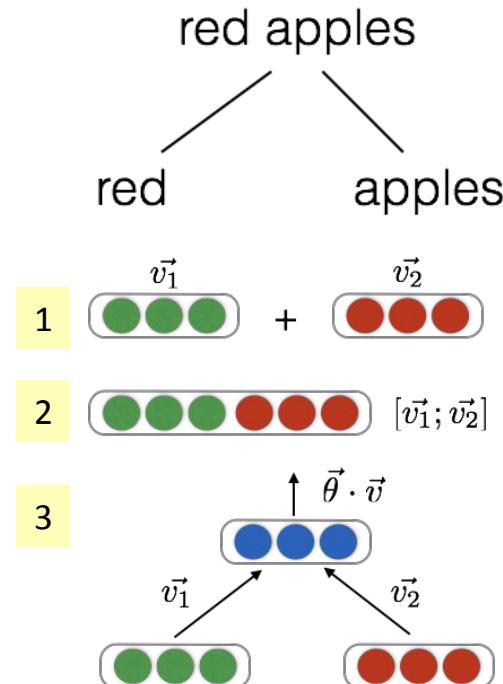
Reasons for exploring Deep Learning

- 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



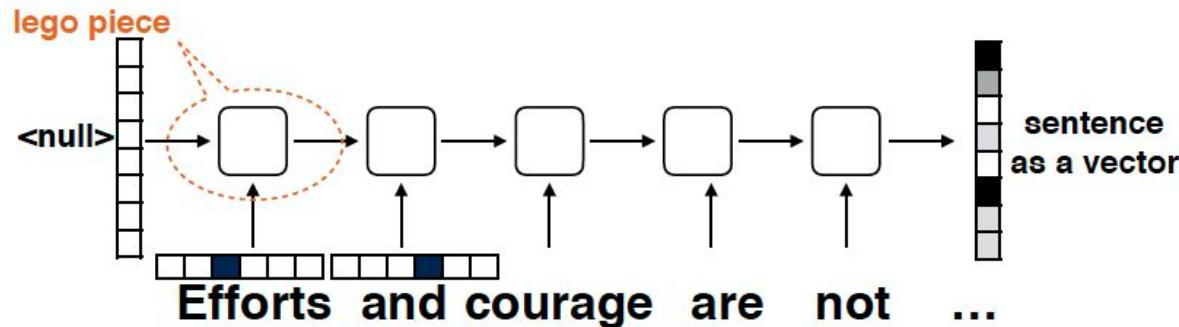
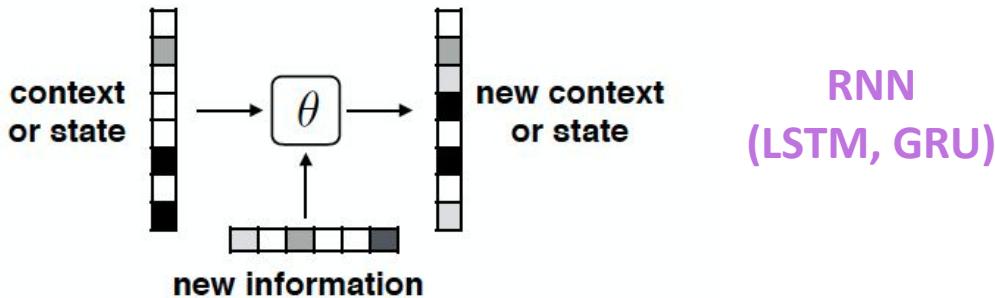
Reasons for exploring Deep Learning (cont.)

- 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



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

หมุกรอบ => |หมู|กรอบ|

ข้าวผัดคนน้ำหมุกรอบหนึ่งจาน => |ข้าวผัด|คน|น้ำ|หมู|กรอบ|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.



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

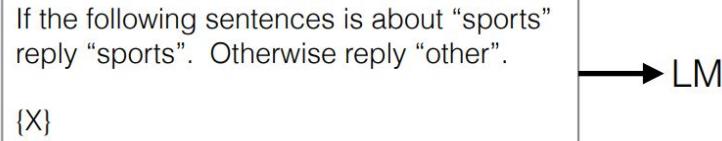
<u>Input X</u>	<u>Output Y</u>	<u>Task</u>
Text	Continuing Text	Language Modeling
Text	Text in Other Language	Translation
Text	Label	Text Classification
Text	Linguistic Structure	Language Analysis
Image	Text	Image Captioning

Methods for Creating NLP Systems

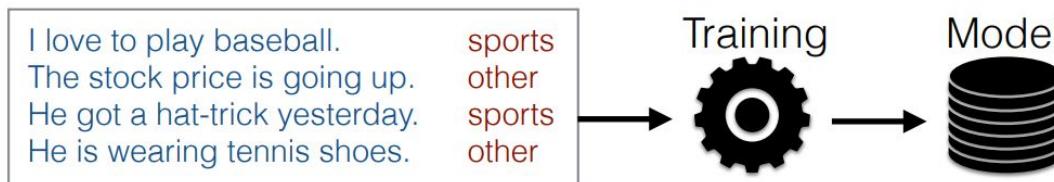
- **Rules:** Manual creation of rules

```
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



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



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)

I hate this movie

positive
neutral
negative

I love this movie

positive
neutral
negative

I saw this movie

positive
neutral
negative

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

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

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

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

Now let's improve!

1. What's going wrong with my system? → 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!

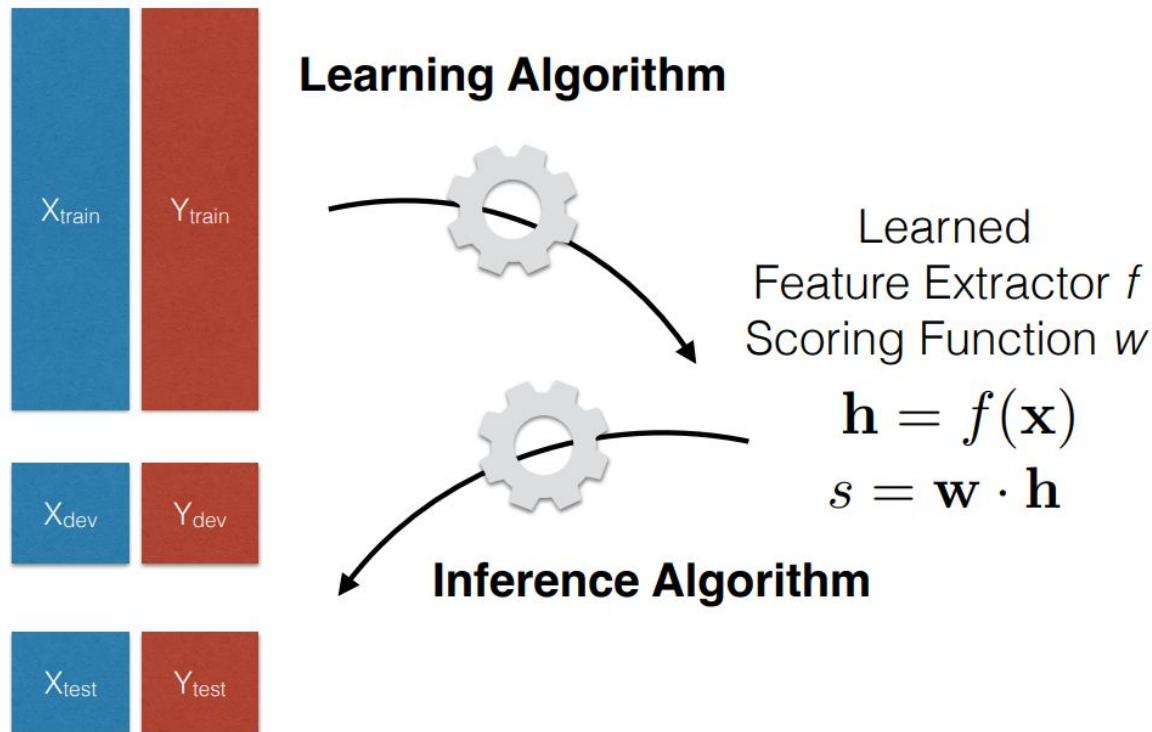
Rule-based Challenges

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? <i>Note: Would require morphological analysis.</i>
Negation	This one is not nearly as dreadful as expected . → positive	If a negation modifies a word, disregard it. <i>Note: Would probably need to do syntactic analysis.</i>
Metaphor, Analogy	Puts a human face on a land most Westerners are unfamiliar with. → positive	???

Demo: Sentiment Classification

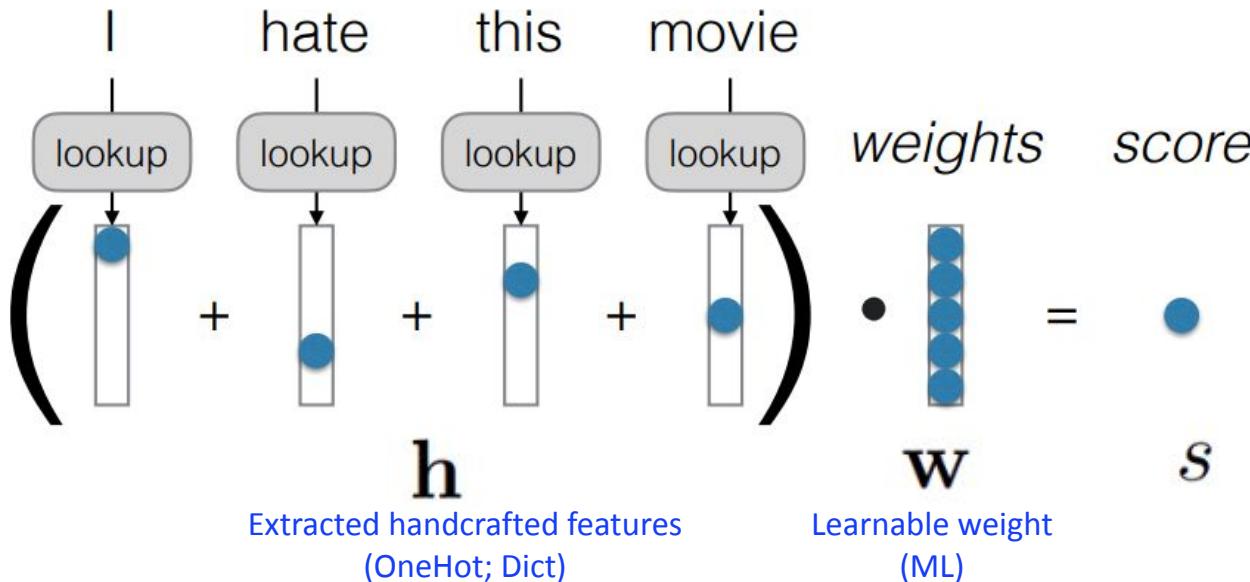
- Rule-Based Model
- **ML-Based Model (BOW)**
- NN-Based Model (not now)

Machine Learning-Based NLP



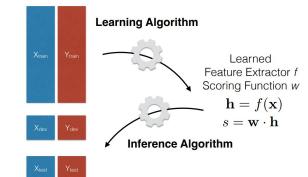
A First Attempt: Bag of Words (BOW)

Aim to solve low freq. words



Features f are based on word identity, weights w learned

Which problems mentioned before would this solve?



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
...	...

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]
            )
```

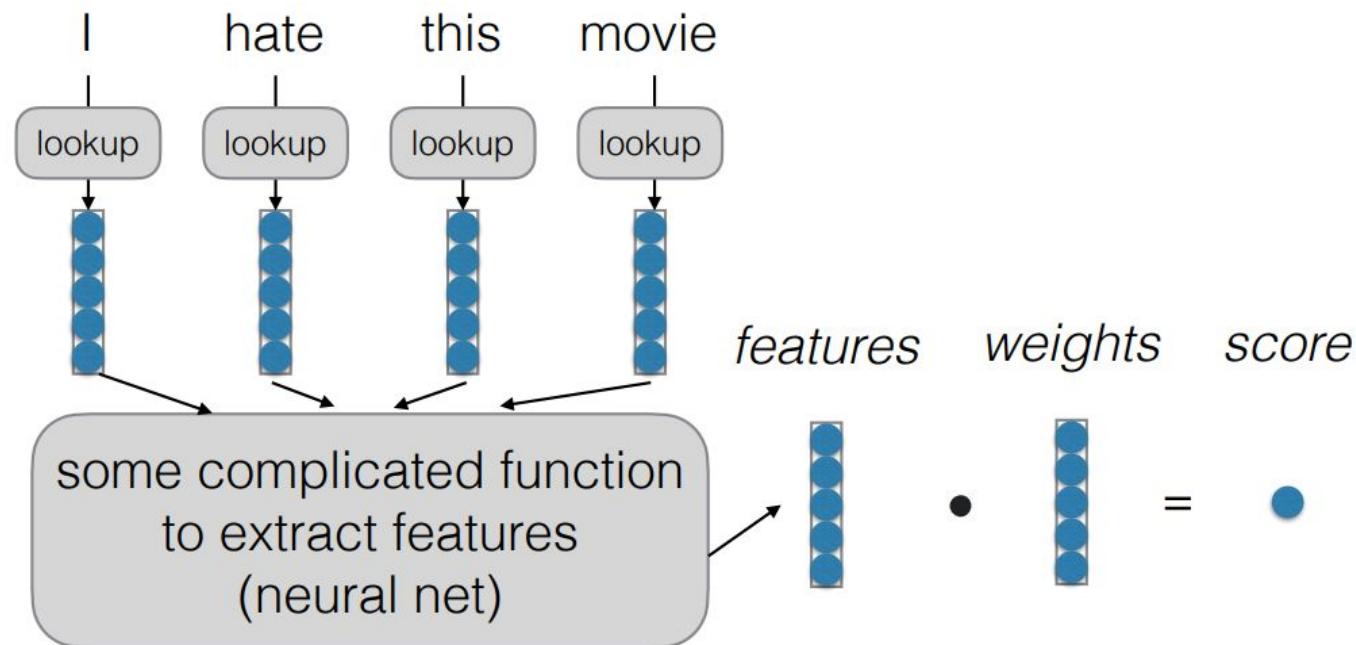
What's Still Missing in BOW?

- 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

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!

4) This Course

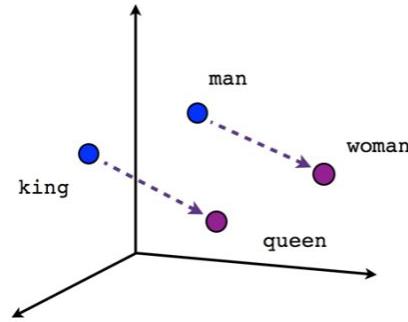
- Thai NLP Challenges
- Class Schedule
- Class Grading

In this class, we ask:

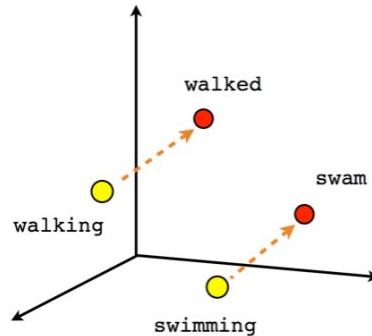
- What goes into the building blocks of **state-of-the-art NLP systems** that **work well** at some tasks?
- Where and why do current state-of-the-art NLP systems still **fail**?
- How can we **utilize NLP in real-world applications** given real-world challenges?

Some NLP Challenges

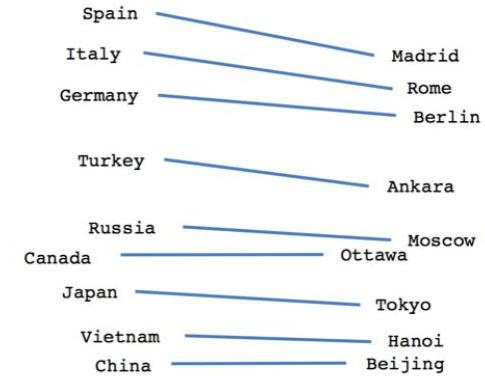
- Complexity in **representing**, learning and using linguistic/situational/world/visual knowledge



Male-Female



Verb tense



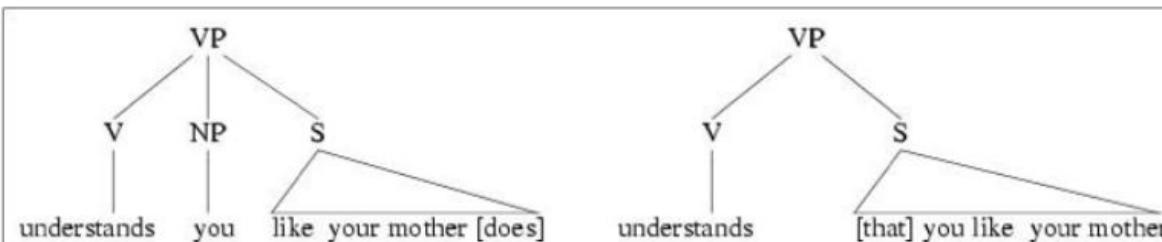
Country-Capital

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]

Thai NLP Challenges

Word segmentation

- **No word delimiters**
- ฉัน|นำ|ดอก|ไม้|ไป|ให้ว|ศาลา|พระ|ภูมิ|ที่|โรง|เรียน|ประจำ|
- ฉัน|นำ|ดอกไม้|ไป|ให้ว|ศาลาพระภูมิ|ที่|โรงเรียน|ประจำ|
- ฉัน|นำ|ดอก|ไม้|ไป|ให้ว|ศาลา|พระภูมิ|ที่|โรง|เรียน|ประจำ|
- ฉัน|นำ|ดอกไม้|ไป|ให้ว|ศาลาพระภูมิ|ที่|โรงเรียนประจำ|

Sentence segmentation

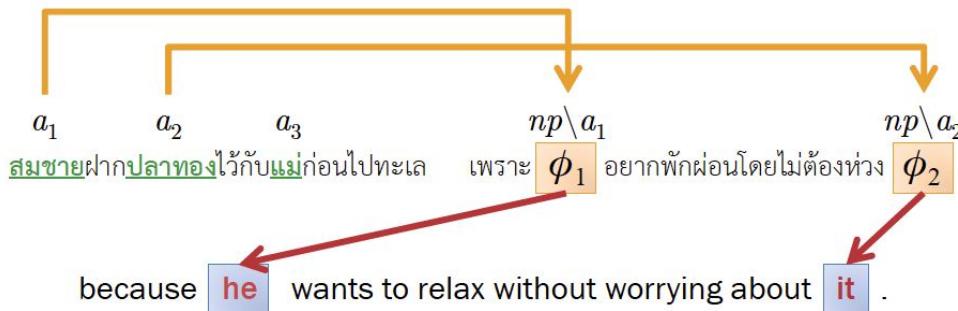
- **No sentence boundary markers**

อย่างไรก็ตาม อดีตประราณ ทปอ. กล่าวว่า มีการหักหัวเรื่องนี้มาตลอดว่า มีช่วงเวลาว่างนานขนาดนี้ ทำไมถึงยังต้องมีการจัดสอบนอกเหนือจากนี้อีก เพราะการสอบล่วงล้าไปในเวลาระหว่างเรียนมั้ยมั่นกระทำบกันเรื่องอื่นๆ โดยเฉพาะการเรียนในชั้นเป็นวงจรลูกโซ่ แนวโน้มที่เข้ามาแก้เรื่องนี้ เป็นความคิดที่ดี แต่ยังไม่เห็นเรื่องใช้ผลการเรียนในชั้นมาเป็นครรภากอบรับตรงซึ่งอาจทำให้เด็กไม่สนใจห้องเรียน และมุ่งความวิชา ทำให้การสอบเข้าอุดมศึกษาตกเป็นจำเลยข้อหาทำลายระบบการศึกษาขั้นพื้นฐาน วนไปสู่ปัญหาเก่าๆ ได้

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

เช้วยร์ป้าดบีนชีนไปเลย โอเคซึ่ง เดี๋ยวคง งานก้า
หุย-แหลก ชั้นกะบวຍ. เอื้อหัวใจเรื่อง เสร็จโต.
บีบีโก ชาไปต้อย สะแಡວแห้ว อู๊ไปเก็บอย เดดะสະมອะ
ໂທລ່ຍໂທຍ ຖັນຮ້ານຮາຄາເກ່າໄຫວ້ ສະມະບະແຮປ ຈິບຈອຍ
ຮັກຄົນບໍອ້າ ແຕ່ລົກໃກ້ໃກ້ພາແນແຮ້ນ ແອັບເປົ້ວລອອື່ນຕົນ ຂອງແກ້ຕ້ອງຈີ 5 ບຸ້ງ
ໄວ້ຖຸກຮົກບໍແລັດ ບາຍຄົດເໜີ້ອນຈັນໄຫມບີ 1 ໄນເຕີມບາກ.
ບ່ອຍໄມ່ຕື່ມຄະ ຄົກບຸ້ ອາໂນແນ ສູນ.ຍົກ. (ສ້າມຍາກ ອ່າງໆຫວຸງ) ເດີສະຫຼື
ຜະຈັນ ທັນ ດັບເປົ້ວລອອື່ນຕົນ ດັບເປົ້ວລອອື່ນຕົນ ເຊິ່ງແລ້ວ
ຫຼັນອົມແບນ ເຕີກຫາວັດ ຕັນ ດັບເປົ້ວລອອື່ນຕົນ ເຊິ່ງແລ້ວ ເຮົຟ
ຕົວຕັ້ງໂທນັ່ງ ໂນເວ ສະເຫຼັບໆ ບັນຫາແທກໝາມໂນ່ມີຮັບຍັບ ໃຫ້ຫາຍເກອະໂຮບິນ
ປະກັບໃຈຈົດ ຈ້ອຍແດກ ກີບເກົ່າເຮົາ ສົຍນິກີຍ
ຄຸນທຸລອກດ້ວຍ ຈ້ອຍແດກ ກີບເກົ່າເຮົາ ສົຍນິກີຍ
ສາຍບ່ອ້ອສ ສະຫຼອບໂນວົວ

ເພື່ອຮັບເອົາຕົວເລີດສະບັບນີ້
www.facebook.com/bungerd2518
IG : bungerd_2518

อัปเดต 60 ศัพท์วัยรุ่น 2567 ตัวตึง รู้ไว้ไม่งงเข้าคุยอะไรกัน

☞ ผู้เขียน : Sale Here Editor

⌚ 12 ก.ย. 67 (วันที่อัปเดตล่าสุด)



[https://salehere.co.th/articles/
update-teen-words](https://salehere.co.th/articles/update-teen-words)

อัปเดต

20 ศัพท์วัยรุ่น ตัวตึง

รู้ไว้ไม่งงเข้าคุยอะไรกัน

คำศัพท์	คำแปล
ตัวตึง	ตัวก็อป , ตัวเต็ง
ตัวแน่	มีความมั่นใจ เช่น เก่ง
บุค	ไม่เริด , ไม่ถูกใจ , ไม่ดี
ความโป๊ะเป็นศูนย์	ไม่มีกึ่ติ
ดุย	หน่อย , ไม่แรง
จ้อชี้	โกรก
เพื่อน	สถานะไม่ชัดเจนระหว่างเพื่อนกับแฟน
ว่าซ่าน	อ่อนว่ากัน , ประมาณนั้น
จะแล้วไหหน	จะเสร็จให้หมด , เสร็จหรือยัง
ฟ้าด	จัดการกันตี , สวย
ตำ	ไปซื้อ
เอาหมัดไม่สนบว่าลูกกิจ	เอาหมัดไม่ว่าของสิ่งนั้นเป็นของใคร
สู๊ด	เพียงมาจาก สู่สุคติ
จัง	ตีมาก , สวย , เริด
มาเอาໄ	มาทำอะไร , มาเอาอะไร
อย่าหาทำ	มาทำอะไร , มาเอาอะไร
หยุดดิ (เสียงสูง)	อย่าทำ , หยุด ไม่ให้ทำ
แบบตะโกน	ตีกันสุด , ตีมากกว่าปกติจนต้องตะโกน

Sale Here

Class Schedule https://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% \times 10 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

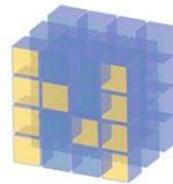
Chapter

Part I: Fundamental Algorithms

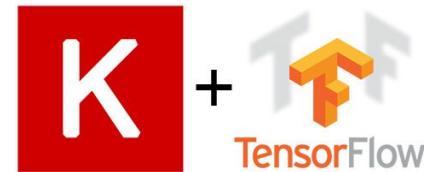
- 1: Introduction
- 2: [Regular Expressions, Tokenization, Edit Distance](#)
- 3: [N-gram Language Models](#)
- 4: [Naive Bayes, Text Classification, and Sentiment](#)
- 5: [Logistic Regression](#)
- 6: [Vector Semantics and Embeddings](#)
- 7: [Neural Networks](#)
- 8: [RNNs and LSTMs](#)
- 9: [Transformers](#)
- 10: [Large Language Models](#)
- 11: [Masked Language Models](#)
- 12: [Model Alignment, Prompting, and In-Context Learning](#)

5) NLP Tools

Implementation



NumPy



PyTorch

PyTorch Lightning

Google
colab

NLP Libraries

NLTK 3.5 documentation

[NEXT](#) | [MODULES](#) | [INDEX](#)

Natural Language Toolkit

nltk 3.9.1

pip install nltk

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.



39,335

text-classification

Search model

token-classification

t: Most downloads ▾

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

The screenshot shows the PyThaiNLP documentation homepage. At the top, there's a header with the PyThaiNLP logo and a search bar. Below the header is a sidebar with links to 'NOTES', 'FAQ', 'Command Line', 'Getting Started', 'Tutorial Notebooks', 'Installation', 'FAQ', 'License', and 'PACKAGE REFERENCE'. The 'PACKAGE REFERENCE' section lists various sub-modules: pythainlp.ancient, pythainlp.augment, pythainlp.benchmarks, pythainlp.chat, pythainlp.classify, pythainlp.coref, pythainlp.corpus, pythainlp.el, pythainlp.generate, pythainlp.khavée, pythainlp.morpheme, pythainlp.parse, pythainlp.phayathaibert, pythainlp.soundex, and 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.khavée

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



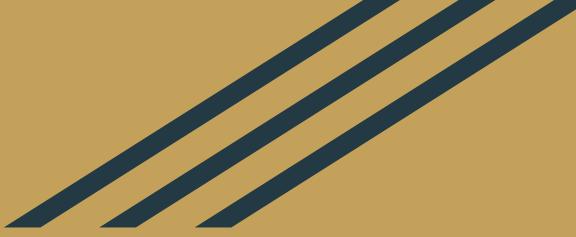
LangChain



LangSmith



LangGraph



Any questions? :)