

# Introduction to Natural Language Processing

## 2110572: NLP SYS

**Peerapon Vateekul & Ekapol Chuangsawanich**

Department of Computer Engineering,  
Faculty of Engineering, Chulalongkorn University



**Peerapon Vateekul, Ph.D.**



**Ekapol Chuangsuwanich, Ph.D.**



**Chawan Piansaddhayanon  
(TA)**



**Patiphan Wongklaew  
(TA)**



**Kao Panboonyuen, Ph.D.  
(TA)**

# Outlines

What is NLP?

NLP Tools

Why NLP is difficult?

Course Logistics

NLP & Text mining

Google Cloud Demo

Deep Learning

# What is NLP?

Definition

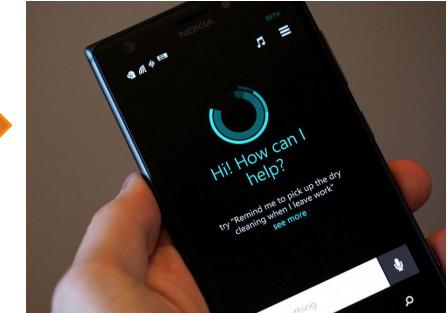
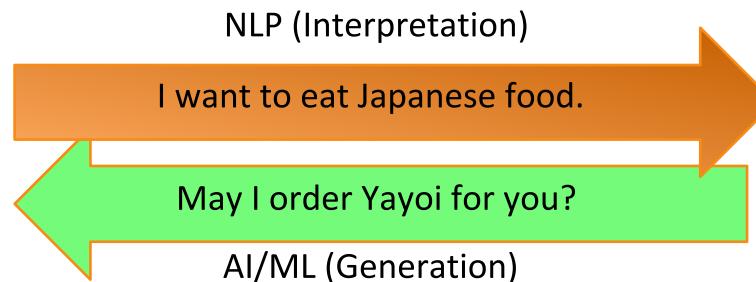
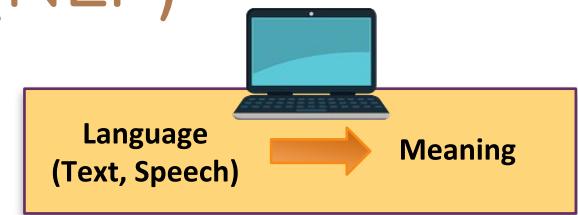
Levels of understanding in NLP

NLP today

# Natural Language Processing (NLP)

NLP is a subfield in AI, where the goal is

- To bridge the gap between **how people communicate** and **what machines understand** in order to perform useful tasks, e.g. making appointments, buying things, question answering, etc.



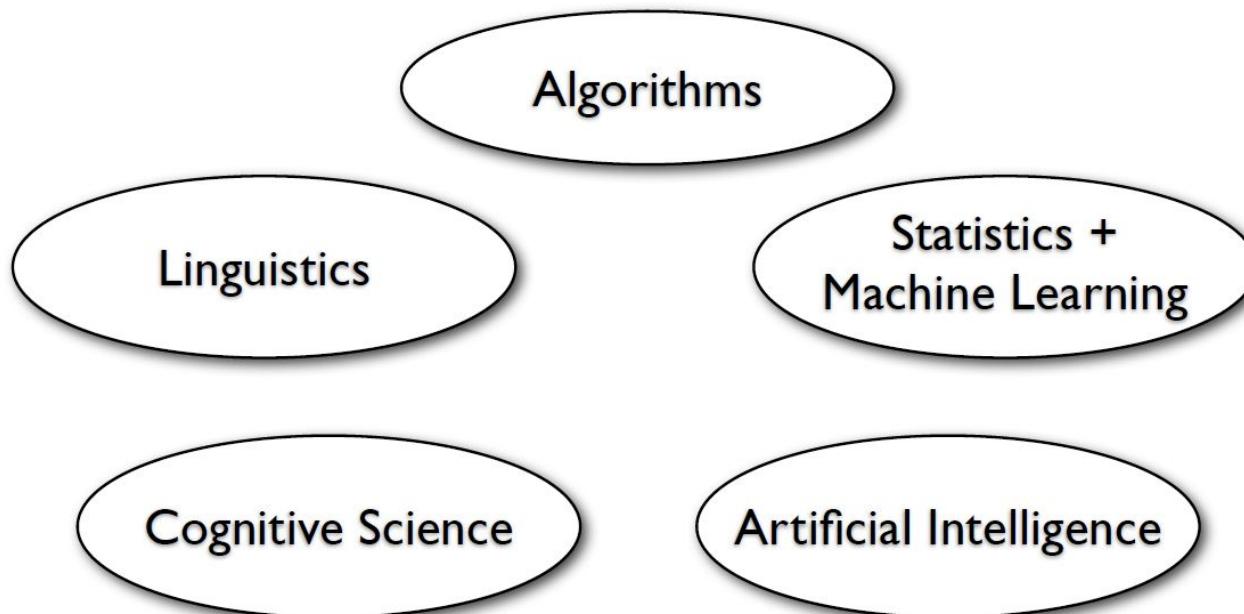
# NLP Goals

Goal: intelligent processing of human languages

- Not just string matching



# NLP is interdisciplinary



# Level of understanding in NLP

[https://www.tutorialspoint.com/artificial\\_intelligence/artificial\\_intelligence\\_natural\\_language\\_processing.htm](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

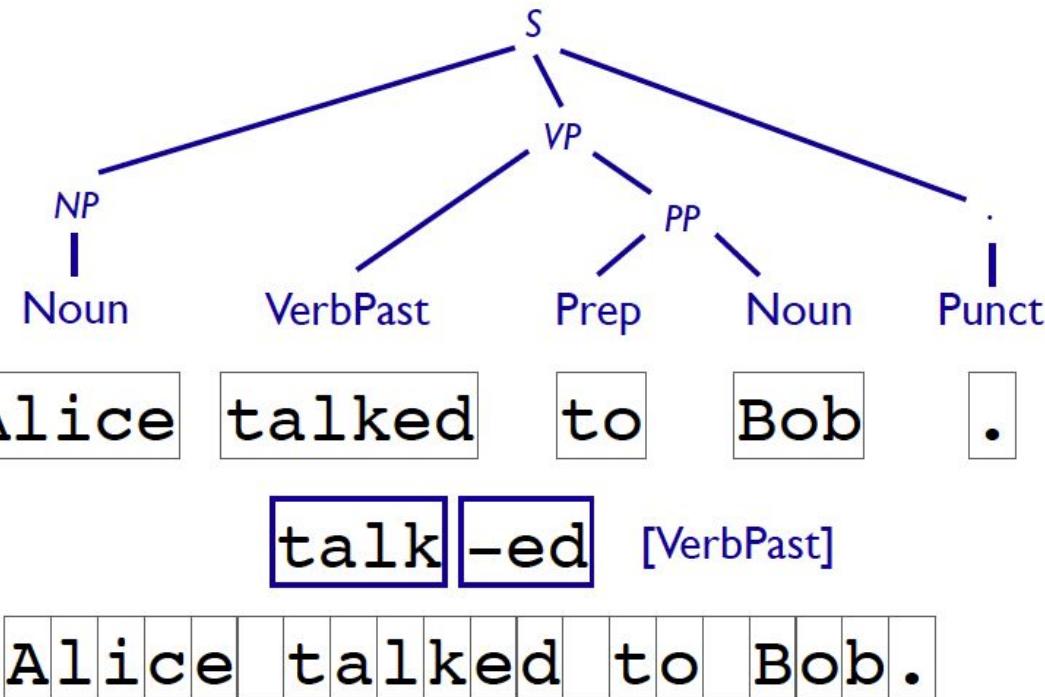
CommunicationEvent(e)

Agent(e, Alice)

Recipient(e, Bob)

SpeakerContext(s)

TemporalBefore(e, s)



# NLP today: Technology

Dan Jurafsky



## Language Technology

making good progress

mostly solved

### Spam detection

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



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

ADJ ADJ NOUN VERB ADV  
Colorless green ideas sleep furiously.

### Named entity recognition (NER)

PERSON ORG LOC  
Einstein met with UN officials in Princeton

### Sentiment analysis

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



### Coreference resolution

Carter told Mubarak he shouldn't run again.

### Word sense disambiguation (WSD)

I need new batteries for my *mouse*.



### Parsing

I can see Alcatraz from the window!

### Machine translation (MT)

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

### Information extraction (IE)

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

still really hard

### Question answering (QA)

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

### Paraphrase

XYZ acquired ABC yesterday  
ABC has been taken over by XYZ

### Summarization

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

### Dialog

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

# NLP today: Machine Translation (MT)

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.

เป็นปีใหม่ที่กำลังได้รับการแสดงความคิดเห็นของผู้เชี่ยวชาญให้บุกมองของพวกเขาก่อนวันสี่ที่ 2018 เป็นในร้าน

ต่อไปนี้เป็นหัวข้อใหญ่สามหัวที่ควรระวังในช่วง 12 เดือนข้างหน้า

การซัมมูนตลาดหุ้นสามารถดำเนินต่อไปได้หรือไม่?

ปีใหม่เริ่มมีตลาดหุ้นในสหรัฐอาณาจักรและสหราชอาณาจักรที่สูงสุดเป็นประวัติการณ์

ดัชนีเฉลี่ยอุตสาหกรรมดาว โจนส์ปรับตัวสูงขึ้นกว่า 25,000 จุดเป็นครั้งแรกในสิบปีที่แล้วที่ดัชนี S & P 500 ที่ใหญ่ขึ้นก็อยู่ในระดับสูงเป็นประวัติการณ์

## Markets, Brexit and Bitcoin: 2018's themes

By Chris Johnston  
Business reporter

5 January 2018

f t m Share



GETTY IMAGES

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

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

# NLP today: Question Answering (QA)



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

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

Ref: Prof. Regina Barzilay, NLP @MIT

# NLP today: Question Answering (QA) (cont.)

บริษัทพาร์ค | ตั้ก, พายไก่, โรเบิร์ต, เปิล | 29 พ.ย. 60 Full HD

<https://www.youtube.com/watch?v=CAJAQUao7HU>



# NLP today: Search/Summarization

Google aquaman

All Images Videos News Maps More Settings Tools

About 164,000,000 results (0.69 seconds)

Showtimes for Aquaman

[Aquaman Movie Official Website - In theaters December 21, 2018](https://www.aquamanmovie.com/)  
https://www.aquamanmovie.com/ ▾  
Aquaman - #AquamanMovie- In theaters December 21st, 2018.

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

Top stories



8:30pm

อควาแมน เจ้าสมุทร

พ.ศ. 2561 · ภาณย์แรมเวฟินคาซี/ภาณย์ครั่วนิยา  
วิทยาศาสตร์ · 2 ชม. 22 นาที

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

94% ชอบภาพยนตร์เรื่องนี้  
จาก Google

อควาแมน เจ้าสมุทร เป็นภาพยนตร์ชูบอร์ด "อควาแมน" ของซีซัมมิกส์ อุตสาหกรรมที่เป็นภาพยนตร์ลีดเดิ้งที่ 6 ในรอบภาพยนตร์จาก จักรราชนิเวศน์ ก้าวโนเบล James Wan ผู้อำนวยการ David Leslie Johnson-McGoldrick และ Will Beall และเรื่องโดย Wan, Beall, และ Geoff Johns นำแสดงโดย Jason Momoa ...

# NLP today: Information Extraction (IE)

## 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  
 ACCESSED 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

# NLP today: IE

## Managing customer data

- JPMorgan Chase has developed [Contract Intelligence \(COIN\)](#) to automate daily routine tasks.
- COIN is used to analyze [documents](#) and extract the important information from it within seconds.
- Saves an estimated 360,000 hours of work each year.
- Actively exploring ways to apply it into other operations.

News > Business > Business News

## JPMorgan software does in seconds what took lawyers 360,000 hours

A new era of automation is now in overdrive as cheap computing power converges with fears of losing customers to startups

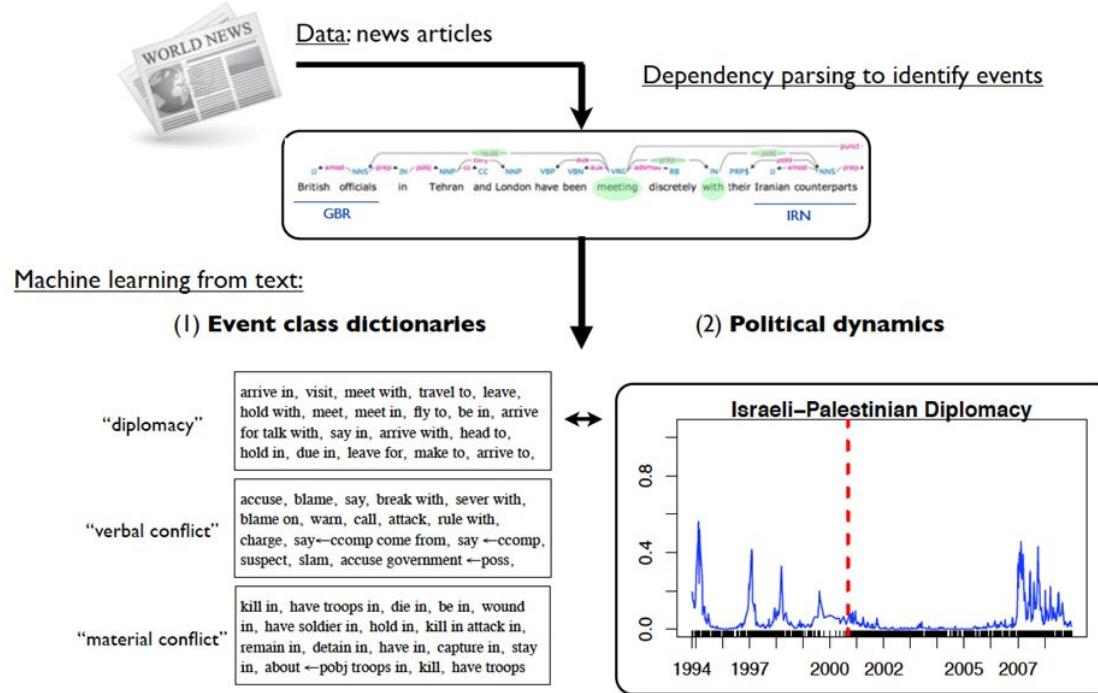
Hugh Son | Tuesday 28 February 2017 11:51 | 16 comments



### JPMORGAN CHASE & Co.



# NLP today: Trend analysis



# 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

BERKSHIRE HATHAWAY INC.  
3555 Farnam Street  
Omaha, NE 68131  
[Official Home Page](#)

- [A Message From Warren E. Buffett](#)
- [Annual & Interim Reports](#)  
Updated November 3, 2017
- [Special Letters From Warren & Charlie RE: Past, Present and Future](#)
- [Link to SEC Filings](#)
- [Links to Berkshire Subsidiary Companies](#)
- [Corporate Governance](#)
- [Owner's Manual](#)
- [Letters from Warren E. Buffett Regarding Pledges to Make Gifts of Berkshire Stock](#)
- [News Releases](#)  
Updated November 3, 2017
- [Warren Buffett's Letters to Be](#)  
Updated February 25, 2017
- [Charlie Munger's Letters to W](#)
- [Annual Meeting Information](#)
- [Celebrating 50 Years of a Pro](#)  
(A commemorative book first sold at the 2015 /
- [Comparative Rights and Relat](#)
- [Berkshire Activewear](#)

GEICO  
FOR A FREE CAR INSURANCE RATE QUOTE THAT COULD SAVE YOU SUBSTANTIAL MONEY  
[WWW.GEICO.COM](#) OR CALL 1-888-395-6349, 24 HOURS A DAY



NLP is difficult!  
Word-level ambiguity!

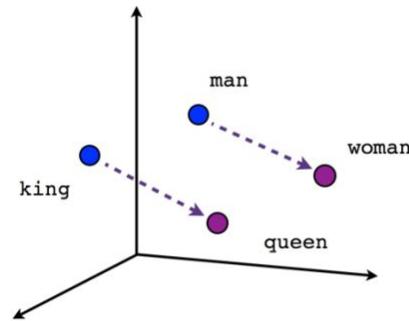
# Why NLP is difficult?

Ambiguity

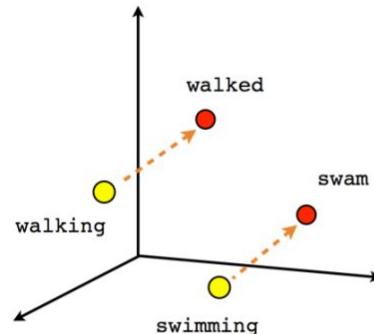
Issues in Thai NLP

# What NLP is difficult?

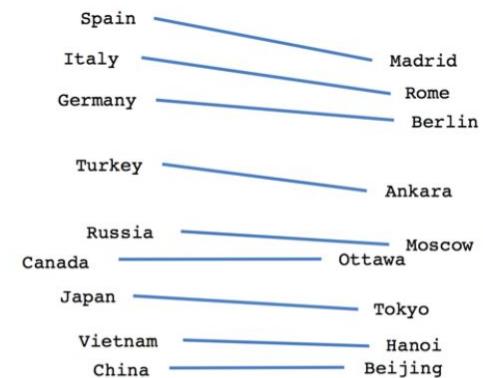
- Complexity in **representing**, learning and using linguistic/situational/world/visual knowledge



Male-Female



Verb tense



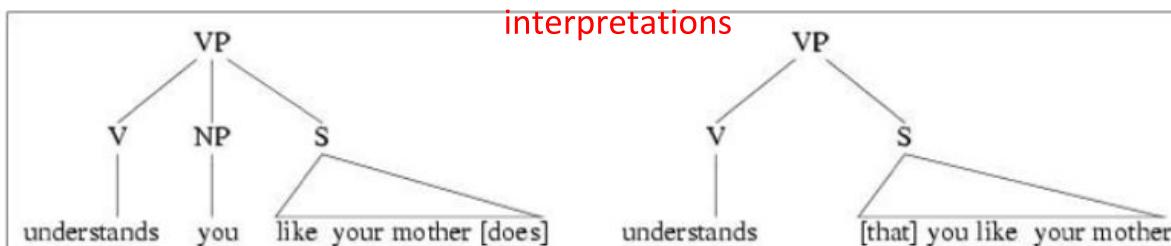
Country-Capital

# What NLP is difficult? (cont.)

- Human languages are **ambiguous** (unlike programming and other formal languages), **so some parts can be ignored**.
- Human languages are interpretation depends 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



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

# Issues in Thai NLP

## Word segmentation

- **No word delimiters**
- ฉัน|นำ|ดอກ|ไม้|ไป|ให้ว|ศาล|พระ|ภูมิ|ที่|โรง|เรียน|ประจำ|
- ฉัน|นำ|ดอກ|ไม้|ไป|ให้ว|ศาล|พระ|ภูมิ|ที่|โรง|เรียน|ประจำ|
- ฉัน|นำ|ดอກ|ไม้|ไป|ให้ว|ศาล|พระ|ภูมิ|ที่|โรง|เรียน|ประจำ|
- ฉัน|นำ|ดอກ|ไม้|ไป|ให้ว|ศาล|พระ|ภูมิ|ที่|โรง|เรียน|ประจำ|

## Sentence segmentation

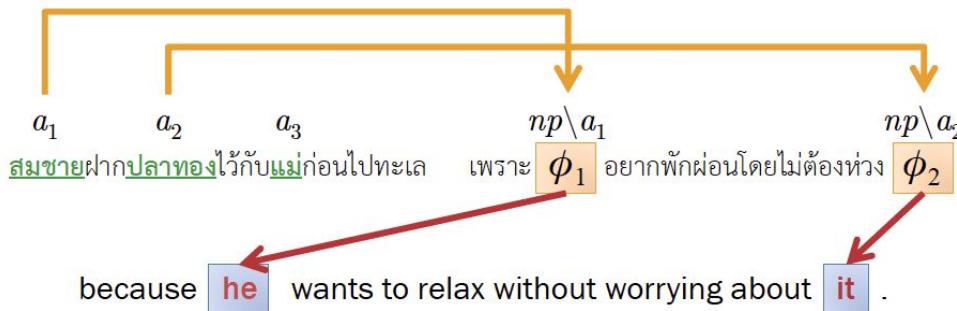
- **No sentence boundary markers**

อย่างไรก็ตาม อดีตประบาน ทปอ. กล่าวว่า มีการหักหัวเรื่องนี้มาตลอดว่า มีช่วงเวลาว่างนานขนาดนี้ ทำไมถึงยังต้องมีการจัดสอบนอกเหนือจากนี้อีก เพราะการสอบล่วงล้ำไปในเวลาระหว่างเรียนมั้ย ยังนั้นกระทบกับเรื่องอื่นๆ โดยเฉพาะการเรียนในชั้นเป็นวงจรลูกโซ่ แนวโน้มที่เข้ามาแก้เรื่องนี้ เป็นความคิดที่ดี แต่ยังไม่เห็นเรื่องใช้ผลการเรียนในชั้นมาเป็นองค์ประกอบรับตรง ซึ่งอาจทำให้เด็กไม่สนใจห้องเรียน และมุ่งกว่าวิชา ทำให้การสอบเข้าอุดมศึกษา ตกเป็นจำเลยข้อหาทำลายระบบการศึกษาขั้นพื้นฐาน วนไปสู่ปัญหาเก่าๆ ได้

# Issues in Thai NLP (cont.)

## Syntax ambiguity

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



## Nostalgic Thai slangs

เช้วยร์ป้าดบีนชีนไปเลย โอเคซึ่ง เดี๋ดดวัง งานก้า  
หุยๆแหนง ชั้งกะบัวย. เชือหัวใจเรื่อง เสร์จโต๊ะ.  
บีบีโก ชาไปต้อย สะడ็ดวงแท้ว อยู่ไปเก็บบอย เดดสัมมาร  
โหลย์โหลย์ ทั้งร้านราคาก่าไหร สะມະນະແゑบ จິບຈອຍ  
รักกันบ້ອງๆ แต่รักกันมากແນ້ນ ਐປັເລືວອົງຕັນ ຂອງແກ້ຕ້ອງຈຸ 5 ບຸນ່າງ  
ໄວ້ຖຸກຮົກບໍແຄືດ ບາຍຄົດເໜ້ນອັນໄຫມບີ 1 ໄນເຕັມບາກ.  
ບ່ອຍໄມ່ຕື່ມຄະ ຄົກບຸ າໂນແນ ສົມ.ຍທ. (ສໝາຍາກ ອ່າງໆຫວຸງ) ເດັດສະຫຼື  
ຈະຈັນ ຫັນອົມແບນ ເຕັກຫາວັດ ຕັນ ຮດເປັນຈາໄຣວະ ເຮົຟ  
ຕັດຕັງໂທນັ່ງ ໂນເວ ສເຕ່ເບຸ້ນ ບ້ານແຕກໜ່າມໂນ່ມຮັບຍັບ ໄຫ້ຕາຍເກອະໂຮບິນ  
ປະກັບໃຈຈົດ ຈ້ອຍແດກ ກີບເກູ່ຍຸເຮົາ ສໍຍົນກີຍ  
ຄຸນທລອກດ້າວ ຈ້ອຍແດກ ກີບເກູ່ຍຸເຮົາ ສໍຍົນກີຍ  
ສາຍບ່ອງອືສ ສະຫຼອບໂນວົວ

ເພື່ອຮັບເຊື້ອຕົວເລີຍ  
www.facebook.com/bungerd2518  
IG : bungerd\_2518



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

# NLP & Text Mining

# NLP & Text Mining

NLP: Language → Meaning

## Text Mining

- Text mining, which is sometimes referred to “text analytics” is one way to make qualitative or “**unstructured data**” **usable by a computer**.
- Convert from unstructured to structured data
- **NLP** techniques are the building blocks for text mining tasks

**NBC Nightly News**  @nbcnightlynews  
America's #1 evening news broadcast.  
Tweets by @newsdel & @braddjaffy. Join us  
on Facebook <http://facebook.com/nbcnightlynews>

Following 

**NBC News**  @NBCNews  
A leading source of global news and  
information for more than 75 years. Have a  
news tip or question? Ask @rozzy,  
@lou\_dubois, @jbaiata or  
@anthonyquintano.

Following 

**CNN Breaking News**  @cnnbrk  
CNN.com is among the world's leaders in  
online news and information delivery.

Following 



Comment	Good	Like	Hate	#
Tweet1	7	8	0	:)
Tweet2	1	0	10	:(
Tweet3	2	9	1	:)

## 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](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)



## 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 ) ]

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

## Named Entity Tags

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

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



## Application

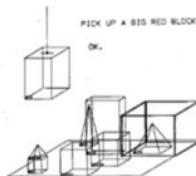
# History of NLP techniques

## 1) Symbolic approach



Noam Chomsky, MIT

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



The dialog that was used as a SHRDLU demo:

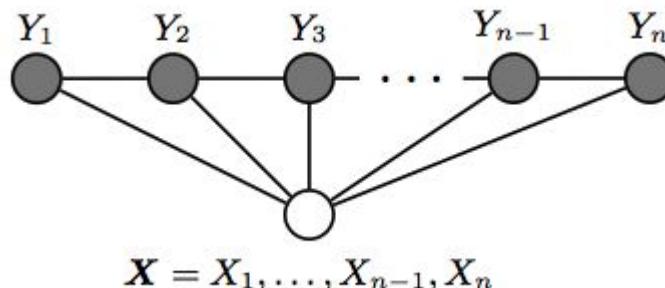
PICK UP A BIG RED BLOCK.  
OK.  
  
Person: PICK UP A BIG RED BLOCK.         
Computer: OK. (does it)  
  
Person: GRASP THE PYRAMID.  
Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.  
Person: FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX.  
Computer: BY "IT", 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?

```
(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))))
```

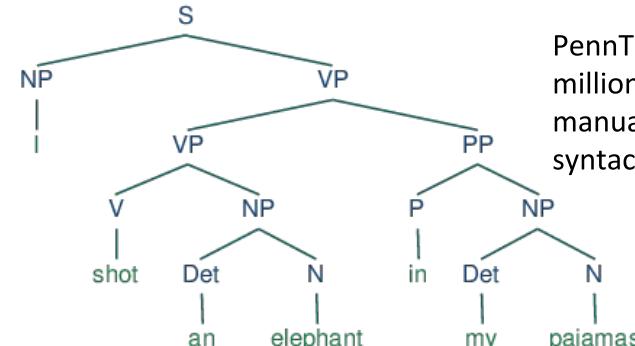
# History of NLP techniques

## 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 is required hand-crafted features.



Conditional Random Fields (CRF)

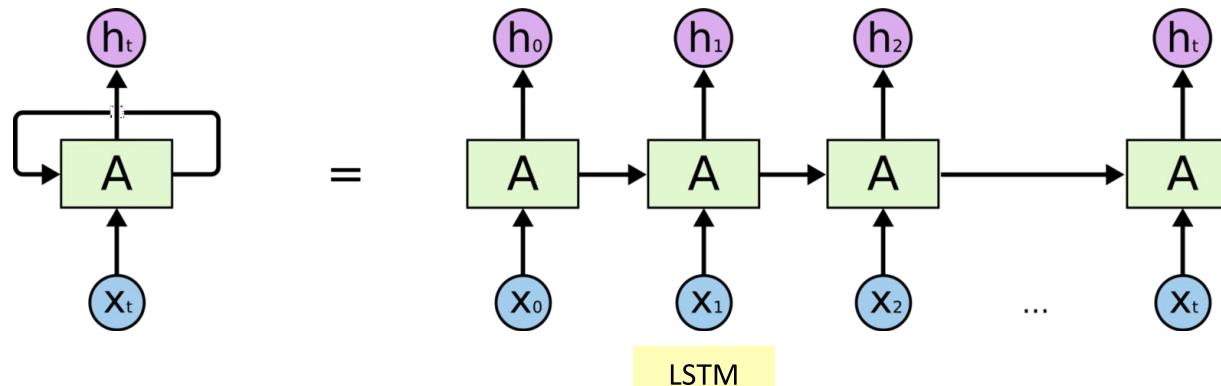


PennTree Bank (1993): one million words from WSJ, manually annotated with syntactic structure

# History of NLP techniques

## 2.5) Deep Learning approach:

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



# Case Study: Determiner placement Symbolic vs. statistical approaches

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

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

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.lmrs2learn.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

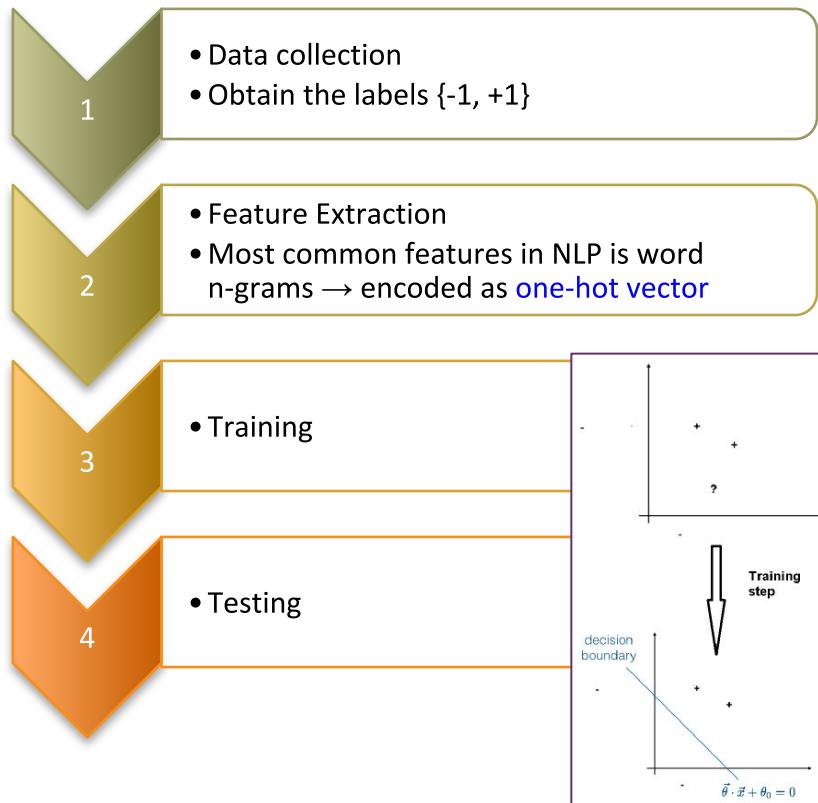
- Consider it as classification
- Predictions: {-1, +1}
- Features:
  - Plural?
  - first appearance in text?
  - head token
  - ...

“lazy monkeys”  
[1 1 0 0 0 … 1]<sup>T</sup>  
↓  
-1

“the United States”  
[1 1 0 0 0 … 0]<sup>T</sup>  
↓  
+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:



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

\*\*\* Self learned features → Deep Learning \*\*\*

*pear*

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

*apple*

↓

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

[0.6 0.2 0.3]

# Deep Learning

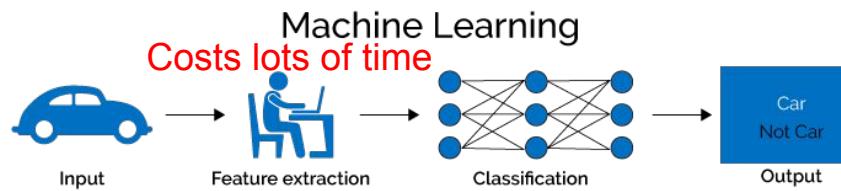
# What is Deep Learning?



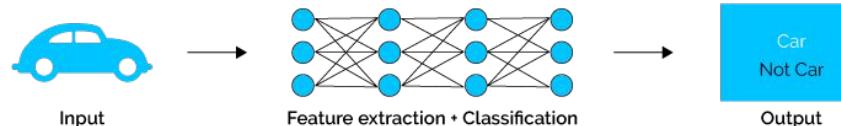
Part of the machine learning field of learning representations of data. Exceptionally effective at learning patterns.



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



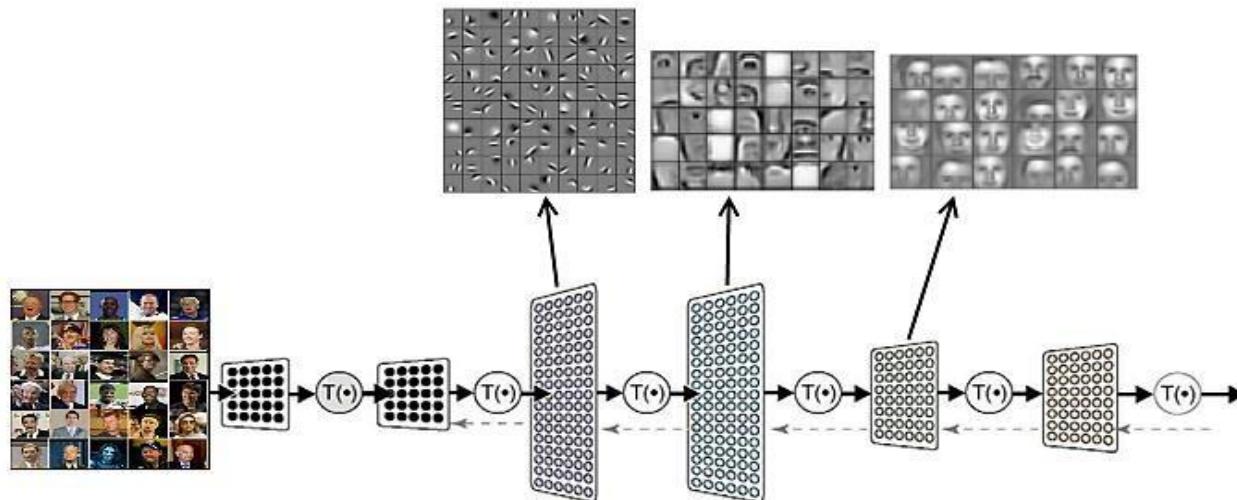
## Deep Learning



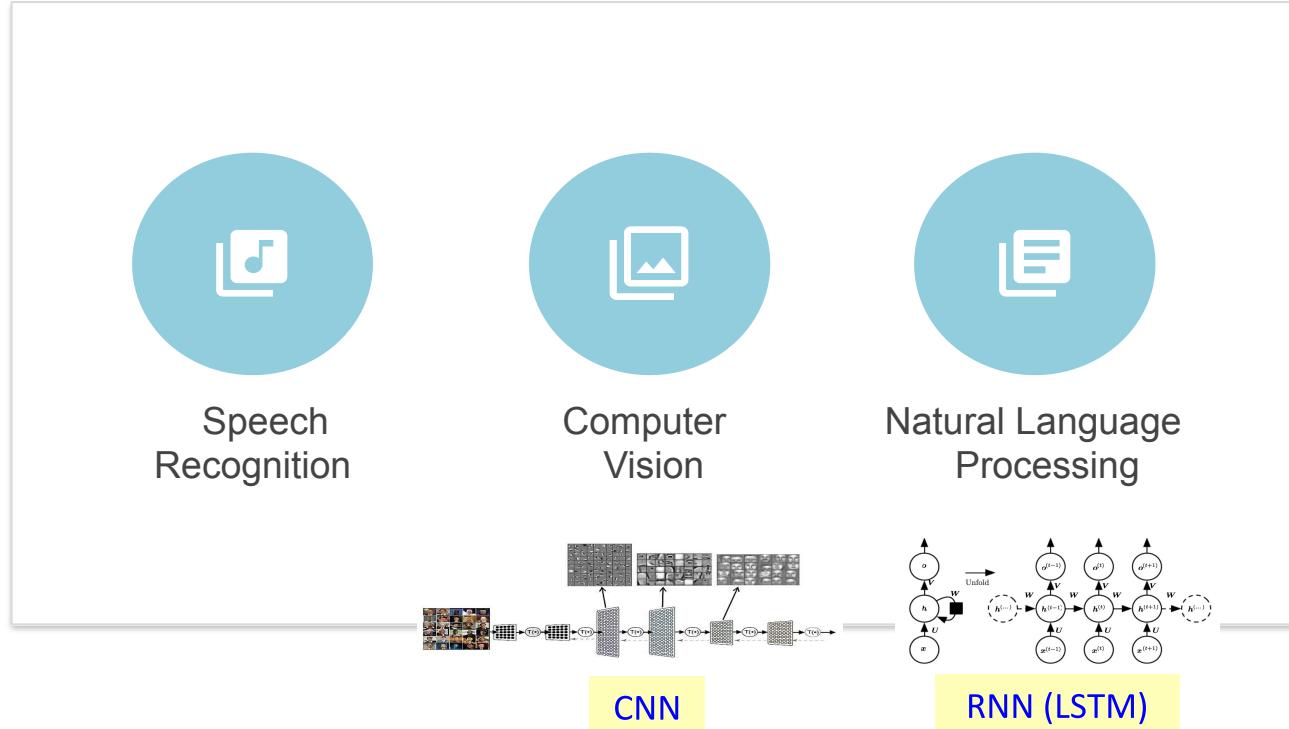
# Deep Learning – Basics (cont.)

## 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)

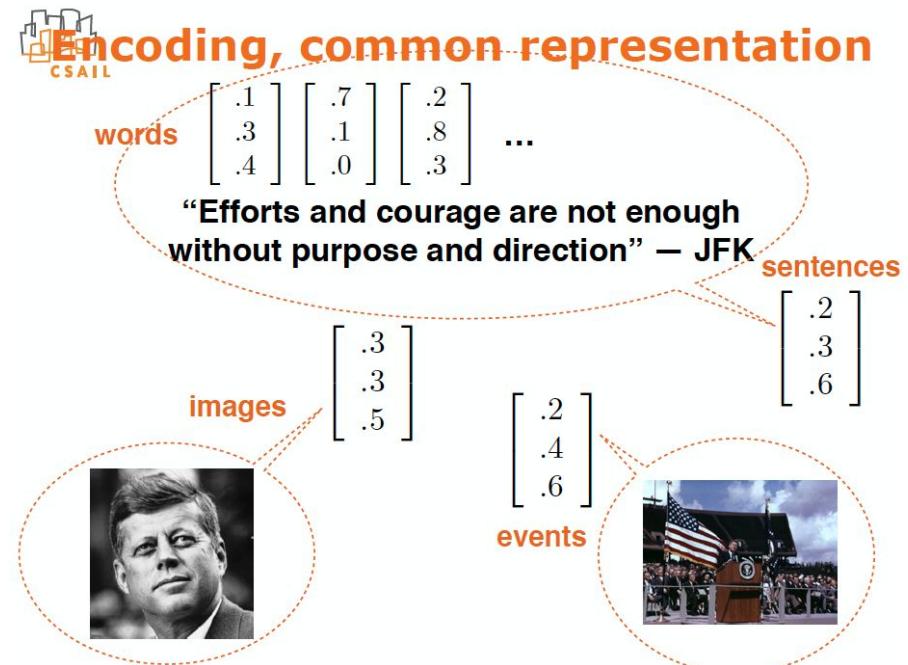
### AttaCut: A Fast and Accurate Neural Thai Word Segmente

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.

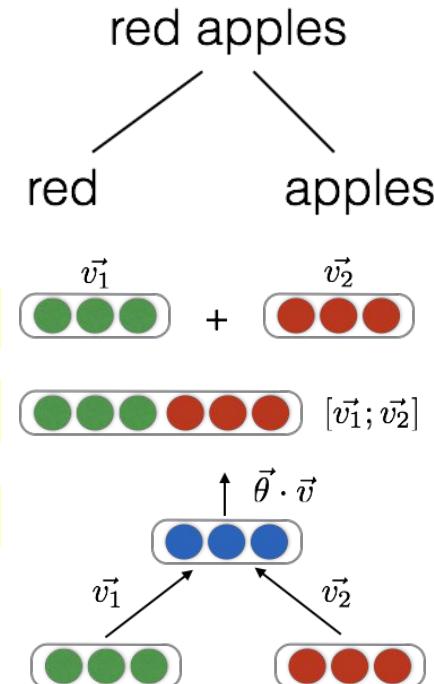
# Reasons for exploring Deep Learning

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



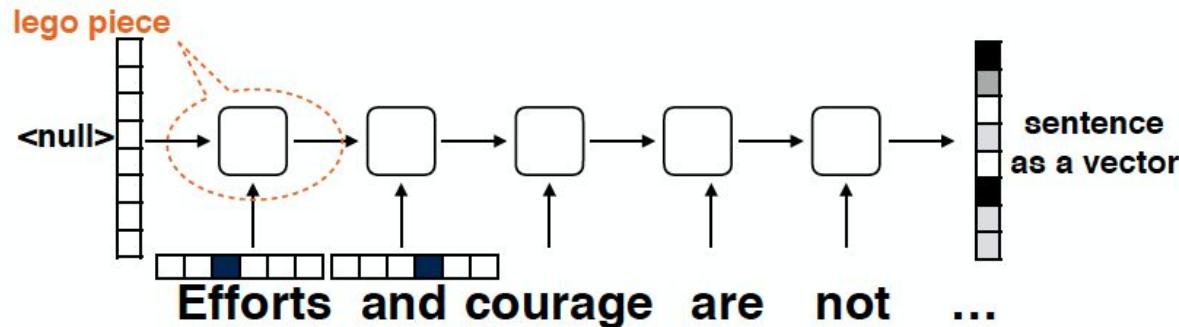
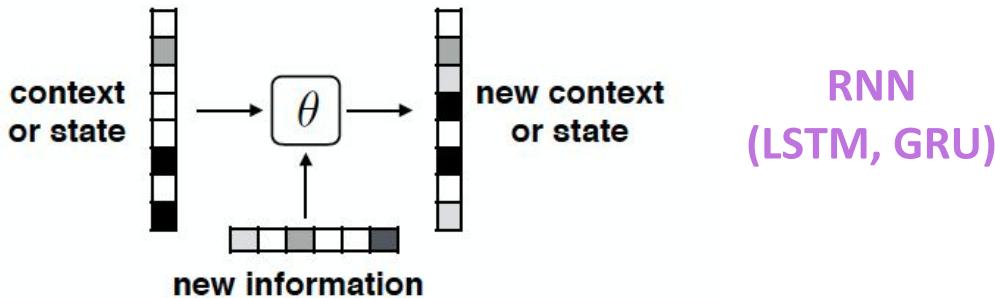
# Reasons for exploring Deep Learning (cont.)

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



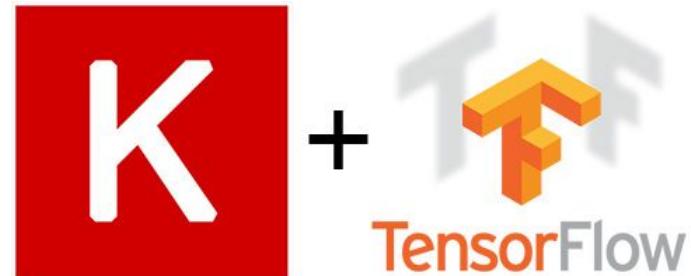
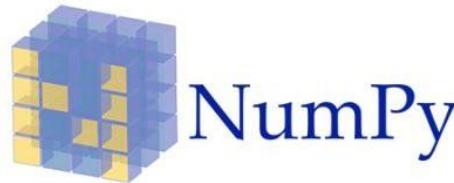
# Reasons for exploring Deep Learning (cont.)

## Example of encoding sentences



# NLP Tools

# Implementation



# NLP Libraries

## NLTK 3.5 documentation

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

## 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.



39,335

text-classification

t: Most downloads ▾

Search model

token-classification

question-answering

multiple-choice

masked-lm

causal-lm

summarization

translation

conversational

zero-shot-classification

ExBERT

<https://thainlp.org/pythainlp/docs/2.0/index.html>

# NLP Libraries for Thai

The screenshot shows the PyThaiNLP documentation homepage. The top navigation bar includes a logo, a search bar, and links for 'NOTES' (Command Line, Getting Started, Installation, From PyThaiNLP 1.7 to PyThaiNLP 2.0), 'PACKAGE REFERENCE' (pythainlp.corpus, pythainlp.soundex, pythainlp.spell, pythainlp.summarize, pythainlp.tag, pythainlp.tokenize, pythainlp.tools, pythainlp.transliterate, pythainlp.ulmfit, pythainlp.util, pythainlp.word\_vector), and a 'Search docs' bar.

Docs » PyThaiNLP documentation

## PyThaiNLP documentation

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

### Notes

- Command Line
- Getting Started
- Installation
- From PyThaiNLP 1.7 to PyThaiNLP 2.0

### Package reference:

- pythainlp.corpus
- pythainlp.soundex
- pythainlp.spell
- pythainlp.summarize
- pythainlp.tag
- pythainlp.tokenize
- pythainlp.tools
- pythainlp.transliterate
- pythainlp.ulmfit
- pythainlp.util
- pythainlp.word\_vector

## 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.

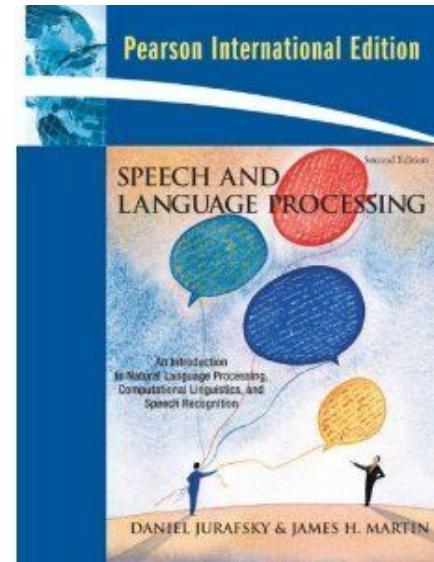
# Course Logistics

# Class Schedule

Mon (9:30-12:30)	Sat (9:00-12:00)	#	Description
18-Jan-2021	23-Jan-2021	1	Intro; Traditional Tokenization (21 first slides)
25-Jan-2021	30-Jan-2021	2	Tokenization
1-Feb-2021	6-Feb-2021	3	PoS Tagging
8-Feb-2021	13-Feb-2021	4	Language Model
15-Feb-2021	20-Feb-2021	5	Word Representation
22-Feb-2021	27-Feb-2021	6	Text Categorization + ประกาย mid-term exam (take-home)
1-Mar-2021	6-Mar-2021	7	Parsing
8-Mar-2021	13-Mar-2021		Midterm Exam Week (8-12 Mar)
			Midterm Exam Submission
15-Mar-2021	20-Mar-2021	8	Attention mechanism & Machine Translation + QA
22-Mar-2021	27-Mar-2021	9	Transformer
			Recent Research in NLP
29-Mar-2021	3-Apr-2021	10	Project Announcement + Paper Announcement
5-Apr-2021	10-Apr-2021	11	NLP Application 1 (Guest); Tentative date: 10-Apr-2021
12-Apr-2021	17-Apr-2021		Songkran Holiday
19-Apr-2021	24-Apr-2021	12	Paper Presentation & Progress Report
26-Apr-2021	1-May-2021	13	NLP Application 2 (Guest); Tentative date: 1-May-2021
3-May-2021	8-May-2021	14	Project Presentation
10-May-2021	15-May-2021		Final Exam Week (10-24 May); No Final Exam

# Course Grading

- Assignments 25%
- Midterm 30%
- Project 45%



## Speech and Language Processing, 2nd Edition 2nd Edition

by [Daniel Jurafsky](#) (Author), [James H. Martin](#) ▾ (Author)

<https://nlp.stanford.edu/~manning/xyzzy/JurafskyMartinEd2book.pdf>



Google Cloud Platform

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

3rd edition draft

**Speech and Language Processing** (3rd ed. draft)

[Dan Jurafsky](#) and [James H. Martin](#)

Here's our December 30, 2020 draft! Includes: