

Introduction to Natural Language Processing

2110594: Natural Language Processing (NLP)

Peerapon Vateekul & Ekapol Chuangsuwanich

Department of Computer Engineering,
Faculty of Engineering, Chulalongkorn University



Peerapon Vateekul, Ph.D.



Ekapol Chuangsuanich, Ph.D.



Can Udomcharoenchaikit
P'Can (TA)



Korrawe Karunratanakul
P'Boat (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

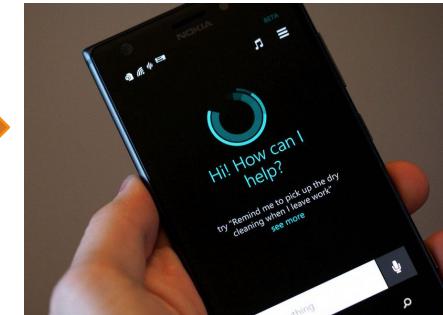
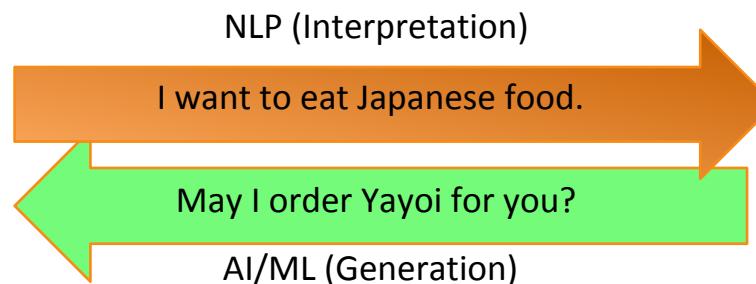
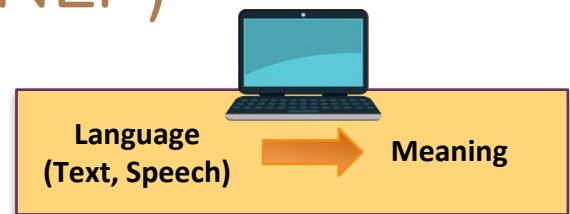
Levels of understanding in NLP

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.



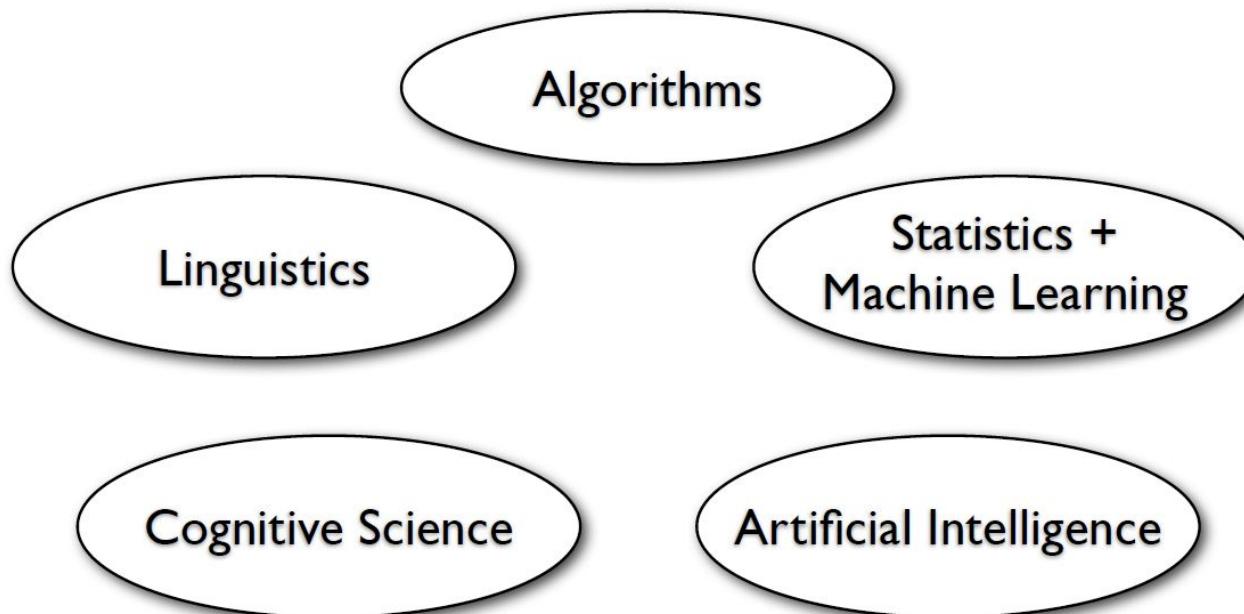
NLP Goals

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

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

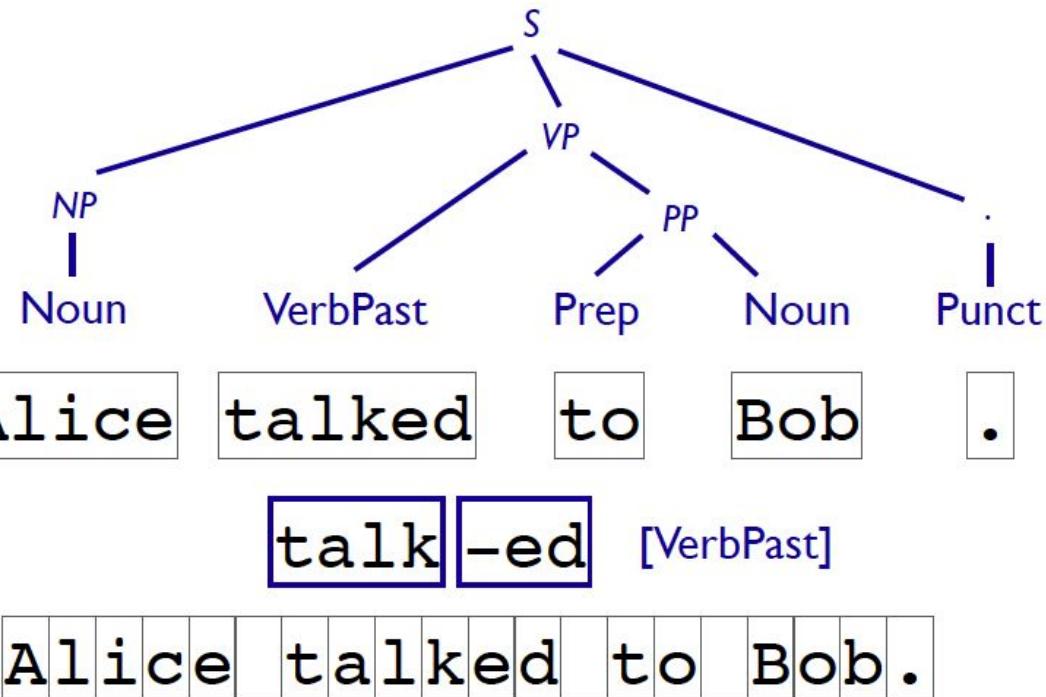
CommunicationEvent(e)

Agent(e, Alice)

Recipient(e, Bob)

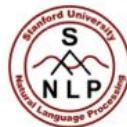
SpeakerContext(s)

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 13th 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.

เป็นปีใหม่ที่กำลังได้รับการแสดงความคิดเห็นของผู้เชี่ยวชาญให้กับมองของพากษาเกี่ยวกับสิ่งที่ 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.)

ปริศนาพ้าแลน | 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 · ภาณย์เรนเวฟน้ำชา/ภาณย์ครั่วนิยา
วิทยาศาสตร์ · 2 ชม. 22 นาที

7.6/10	64%	55%
IMDb	Rotten Tomatoes	Metacritic

94% ชอบภาพยนตร์เรื่องนี้
จาก Google

อควาแมน เจ้าสมุทร เป็นภาพยนตร์ชูบอร์ด "อควาแมน" ของซีซั่นมาสเตอร์ ที่สร้างโดยผู้เขียนบทและผู้กำกับ James Wan และได้รับการสนับสนุนจากผู้อำนวยการ撮影 David Leslie Johnson-McGoldrick และ Will Beall และเรื่องโดย 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	6/12/06	BIOBX-L	0	P		3	CALCS	N
32875	56	1	7/11/06	BIOBX-B	0	N		3		
2247	72	1	4/12/06	BIOBX-R	0	N		3		
45521	61	1	3/30/06	BIOBX-B	0	B		3	CALCS	S
48987	41	1	4/5/06	BIOBX-B	0	P		3	CALCS	N
4179	67	1	5/12/06	BIOBX-B	0	P		2	CALCS	N
24300	59	1	3/31/06	BIOBX-L	0	N		3		
67960	64	1	4/7/06	BIOBX-R	0	P		3	MASS	O
43283	61	W	7/21/06	BIOBX-B	0	B		3		
43319	51	1	4/7/06	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

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

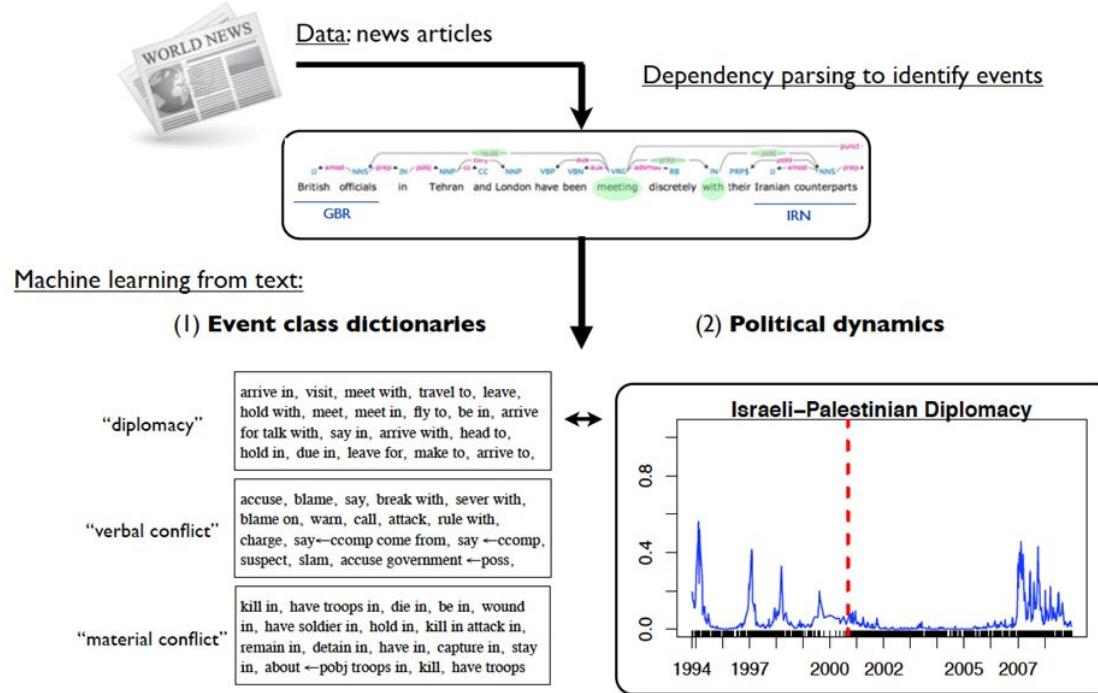
Hugh Son | Tuesday 28 February 2017 11:51 | 16 comments



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!

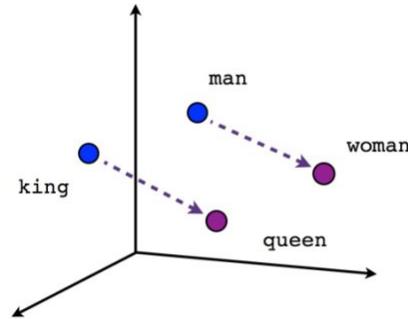
Why NLP is difficult?

Ambiguity

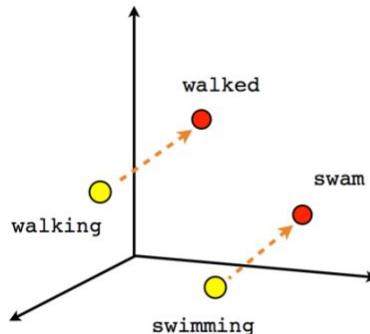
Issues in Thai NLP

What NLP is difficult?

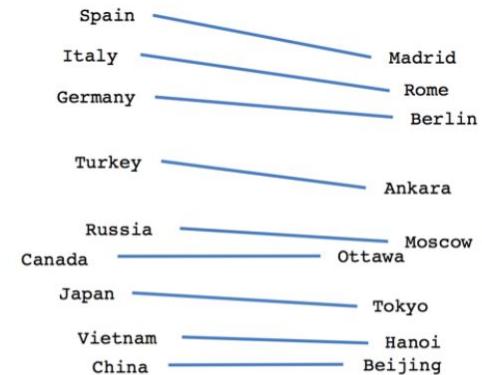
- Complexity in **representing**, learning and using linguistic/situational/world/visual knowledge



Male-Female



Verb tense



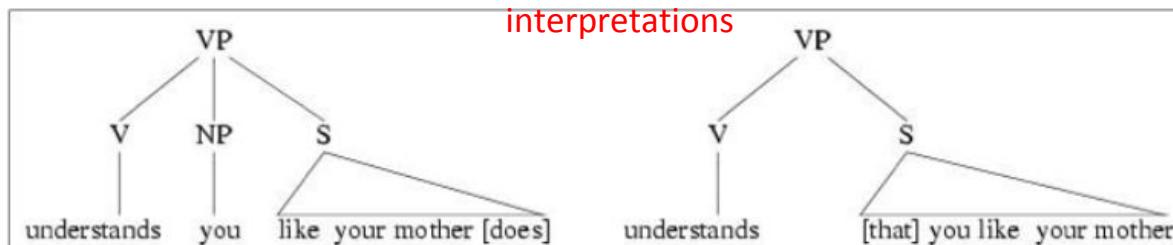
Country-Capital

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

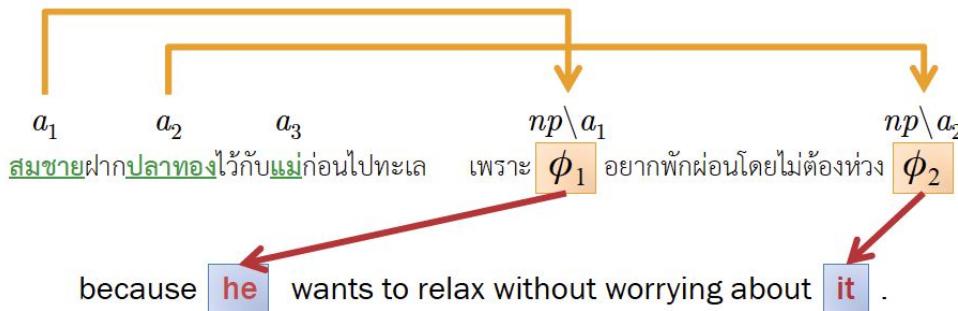
- **No sentence boundary markers**

อย่างไรก็ตาม อดีตประบาน ทปอ. กล่าวว่า มีการหักหัวเรื่องนี้มาตลอดว่า มีช่วงเวลาว่างนานขนาดนี้ ทำไมถึงยังต้องมีการจัดสอบนอกเหนือจากนี้อีก เพราะการสอบล่วงล้ำไปในเวลาระหว่างเรียนมั้ยมั่นนั้นกระทบกับเรื่องอื่นๆ 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

เชัวร์ป้าดบีน เช่น ไปเลย โอเคซึ่ง เดี๋ดดวง งานก้า
หุย แหลก ชั้นกะบวย ใช่หัวใจเรื่อง เสร์จโก^๔
บีโก ชาไปต้อย สะแគแหวว อยู่ไปเก็บอย เดดສະมອะ
ໂທລ່ຍໂທຍ ຖັນຮ້ານຮາຄາເກ່າໄຫວ ສະມະບະແຮປ ຈິບຈອຍ
ຮັກຄົນບໍອ້າ ແຕ່ລົກໃຫ້ພາແນແຮນີ້ ແອັບເປົວອື່ນຕົນ ຂອງແກ້ຕ້ອງໆ 5 ບຸ້ງ
ໄວ້ຖຸກຮົກບໍແຄືດ ບາຍຄົດເໜີ້ອນຈັນໄຫມບີ 1 ໄນເຕີມບາກ
ບ່ອຍໄມ້ດື່ມຄະ ຄົກບຸ ອາໂນແນ ສົມ.ຍທ. (ສໝາຍາກ ອ່າງ່າວວ) ເດີສະຫຼື
ຈະຈັບ ທັນອົມແບນ ເຕີກຫາວົດ ຕັນ ຮດເປັນຈາໄຣວະ ເຮົຟ
ຕະຕັ້ງໂທນັ່ງ ໂນເວ ສະເຫຼັບໆ ບ້ານແຕກໜໍມອ່ນຮັບຍັບ ໃຫ້ຕາຍເກອະໂຮບິນ
ປະກັບໃຈຈົດ ຈ້ອຍແດກ ກີບເກູ່ຍຸເຮົາ ສົຍົນກີຍ
ຄຸນທຸລອກດ້ວຍ ຈ້ອຍແດກ ກີບເກູ່ຍຸເຮົາ ສົຍົນກີຍ
ສາຍບ່ອ້ອສ ສະຫຼອບໂວວ

ເພື່ອຮັບເຊື້ອ ເຊື້ອສະບັບນີ້
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/Fall2003/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)

A word cloud visualization where the size of each word corresponds to its frequency in the input text. The words are colored according to their part of speech tags: PER (red), LOC (blue), ORG (green), and O (black). The most prominent words are 'India' (red), 'China' (blue), 'United States' (green), and 'Nelson Mandela' (black).

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

Application

Tokenization

- Input: ขสมก. เลี้ง จัดหารรถ
- Output: ข, สมก., เล, ีง, จ, ัด, ห, า, ร, ถ

Part of Speech tagging

- Input: [ข, สมก., เล, ีง, จ, ัด, ห, า, ร, ถ]
- Output: [(ข, สมก., NR), (เล, ีง, VV), (จ, ัด, ห, า, VV), (ร, ถ, 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: [(ข, สมก., NR), (เล, ีง, VV), (จ, ัด, ห, า, VV), (ร, ถ, NN)]
- Output: [(ข, สมก., NR, ORG), (เล, ีง, VV, O), (จ, ัด, ห, า, VV, O), (ร, ถ, NN, O)]

Named Entity Tags

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

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



Application

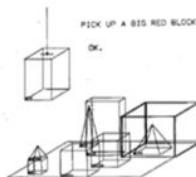
History of NLP techniques

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:

Person: PICK UP A BIG RED BLOCK.
Computer: OK.
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?

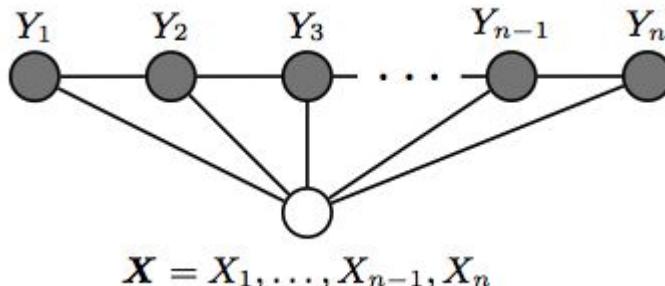
```
(DEFTHEOREM TC-GRASP
  (THCONSE (X Y)  (#GRASP $?X)
            (THGOAL (#MANIP $?X))
            (THCOND ((THGOAL (#GRASPING $?X))
                      ((THGOAL (#GRASPING $_Y))
                       (THGOAL (#GET-RID-OF $?Y)
                           (THUSE TC-GET-RID-OF))))
               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

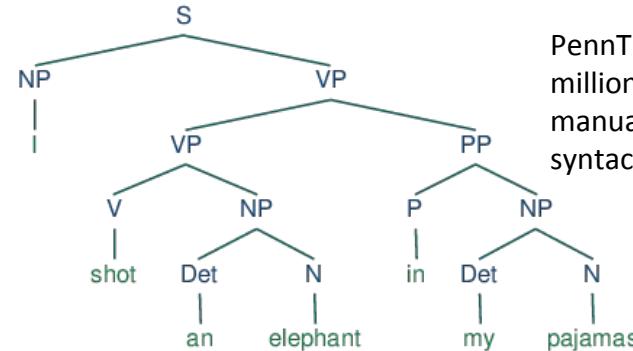
History of NLP techniques

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)

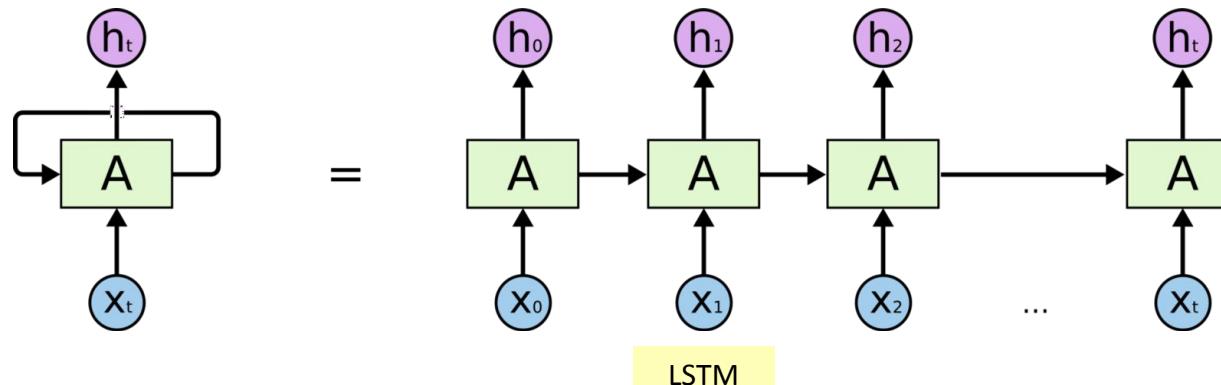


PennTree Bank (1993): one million words from WSJ, manually annotated with syntactic structure

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

www.lmns2learn.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

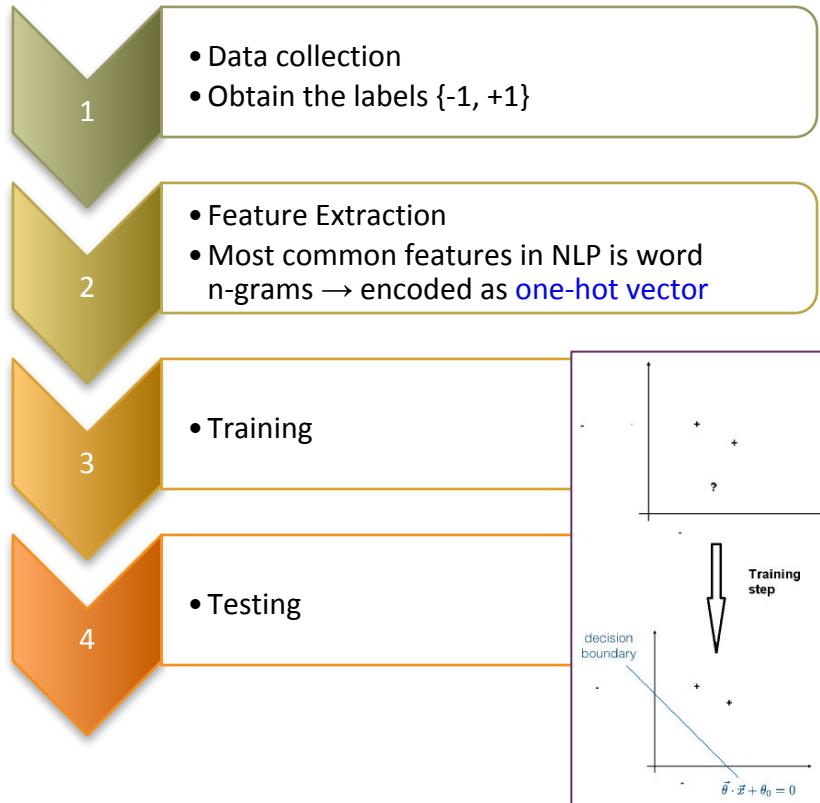
- Consider it as classification
- Predictions: {-1, +1}
- Features:
 - Plural?
 - first appearance in text?
 - head token
 - ...

“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 Map **discrete**, one-hot vectors into low-dimensional **continuous** representations.
***** Self learned features → Deep Learning *****



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

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

pear

[1 0 0 0 ... 0]

apple

[0 0 1 0 ... 0]

↓
[0.4 0.1 0.1]

↓
[0.6 0.2 0.3]

Deep Learning

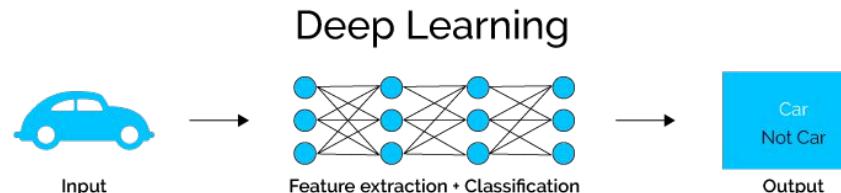
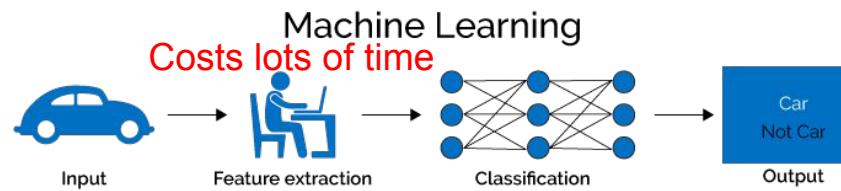
What is Deep Learning?



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



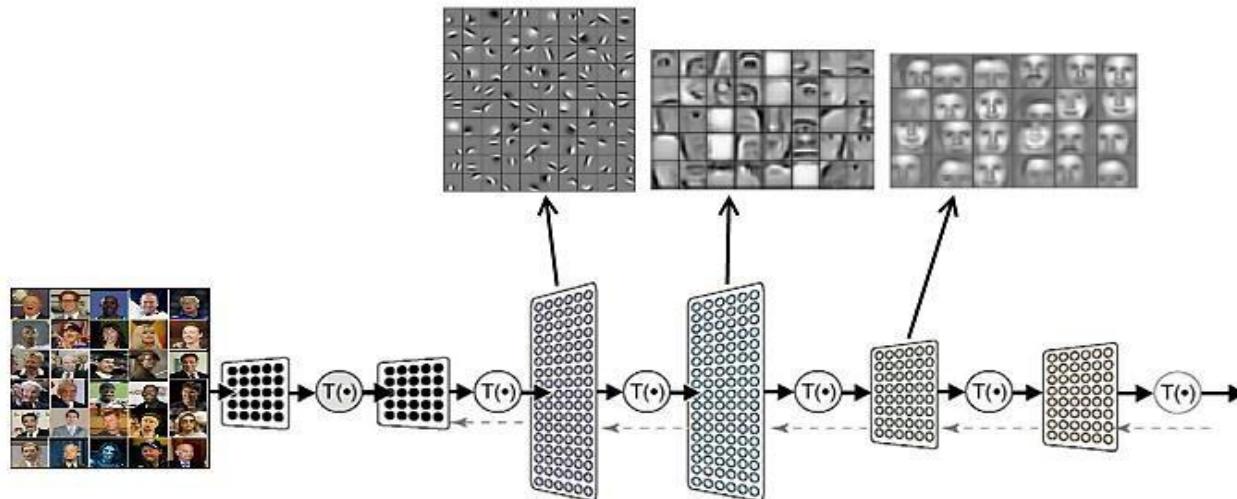
Utilizes learning algorithms that derive meaning out of data by using a hierarchy of multiple layers that mimic the neural networks of our brain.



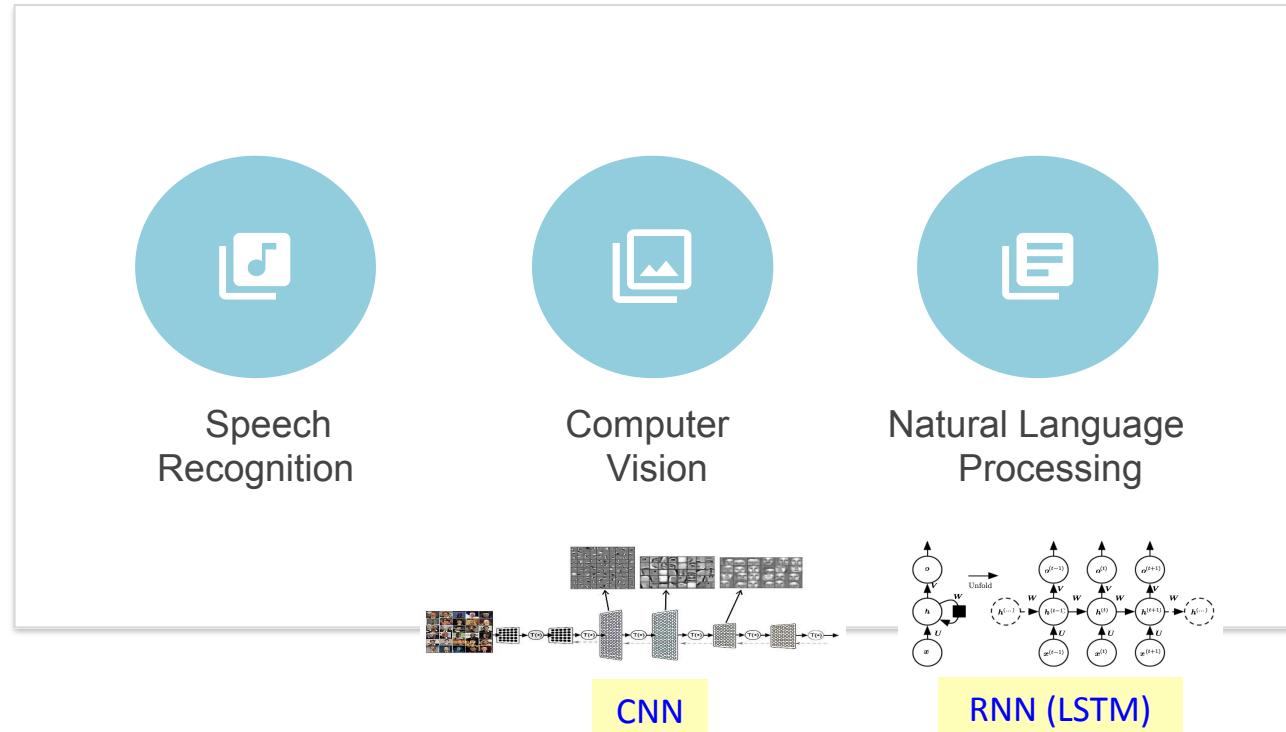
Deep Learning – Basics (cont.)

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



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



pucktada/cutkum
cutkum - Thai Word-Segmentation with Deep Learning in Tensorflow
GITHUB.COM

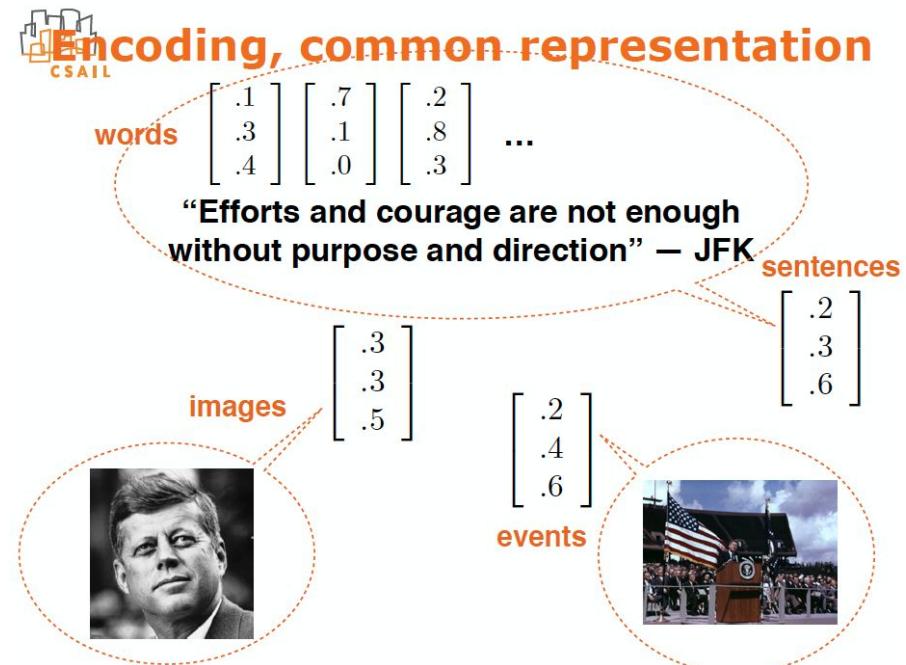
Pucktada Treeratpituk
RNN, F1 = 0.93 on BEST2010

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.

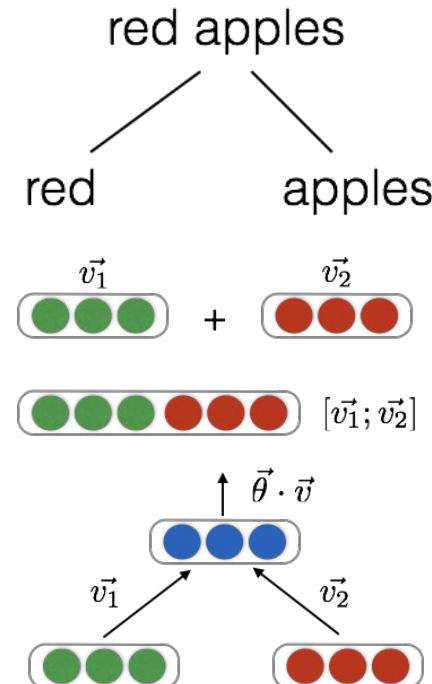
Reasons for exploring Deep Learning

- 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



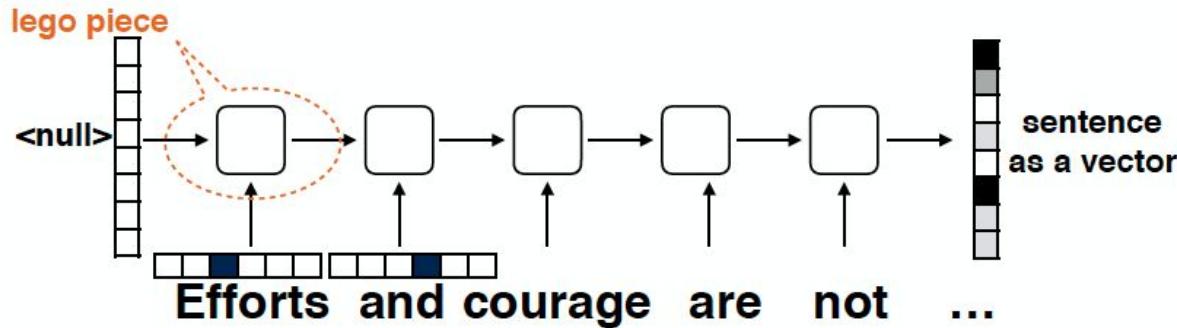
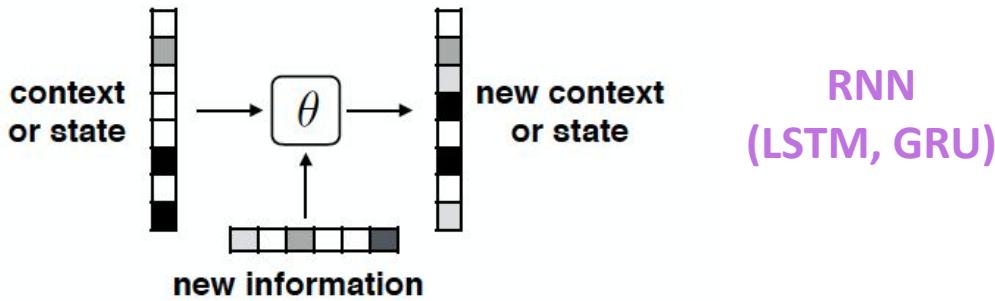
Reasons for exploring Deep Learning (cont.)

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

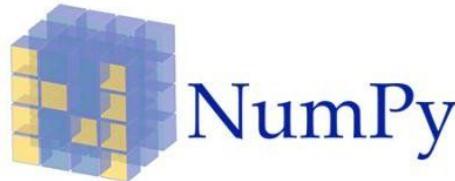


Reasons for exploring Deep Learning (cont.)

Example of encoding sentences



NLP Tools



- <http://nltk.org>
- Tokenization (Parser)
- Stop words removal
- Stemming
- N-gram
- Part-of-Speech Tagging
- Named Entity Recognition
- Etc.

Tools for Thai NLP

Open source

- Word Segmentation
 - CutKum
 - LexTo (NECTEC)
 - SWATH (CMU) – 80's
 - ICU (IBM) – 90's
- POS tagging
 - SWATH (CMU)
- Pythai NLP

Commercial

- Word Segmentation
 - SegIt (NECTEC)
- POS tagging
 - PosIt (NECTEC)

PyThaiNLP - โมดูล NLP ภาษาไทยใน Python

สวัสดีผู้อ่านทุกท่านครับ นหความนี้จะพาผู้อ่านไปใช้งาน PyThaiNLP กันครับ

โนดูล PyThaiNLP เป็นโนดูลที่ถูกพัฒนาขึ้นเพื่องานวิจัยและพัฒนาการประมวลภาษาธรรมชาติภาษาไทยในภาษา Python ที่ดูเหมือนอย่าง ภาษา วรรณ仇恨ช์ ลักษณะไทยบุลป์

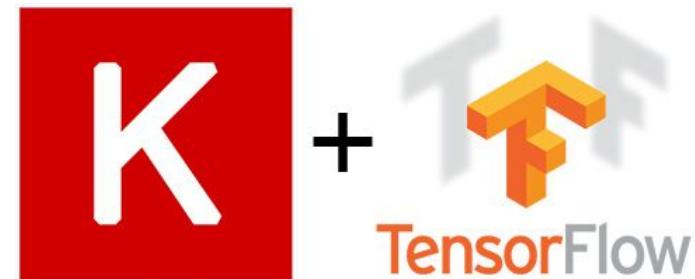
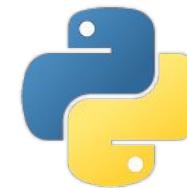
- รองรับทั้ง Python 2.7 และ Python 3.4+
- ใช้ Apache Software License 2.0

โดยมีจุดมุ่งมีความสามารถดังนี้

- ตัดคำภาษาไทย
- ลดเสียงภาษาไทยเป็น Latin
- Postaggers ภาษาไทย
- WordNet ภาษาไทย
- อ่านตัวเลขเป็นมือความภาษาไทย
- เรียงลำดับคำของประโยค
- แยกเป็นคุณภาพที่มีคุณภาพสูง
- เช็คคำศัพท์ในภาษาไทย
- และอื่นๆ

Deep Learning tools

- Tensorflow (Google): Python, etc.
- Keras (François Chollet): Python
- Torch (The Idiap Research Institute):
 - PyTorch: Python



Course Logistics

Class Schedule

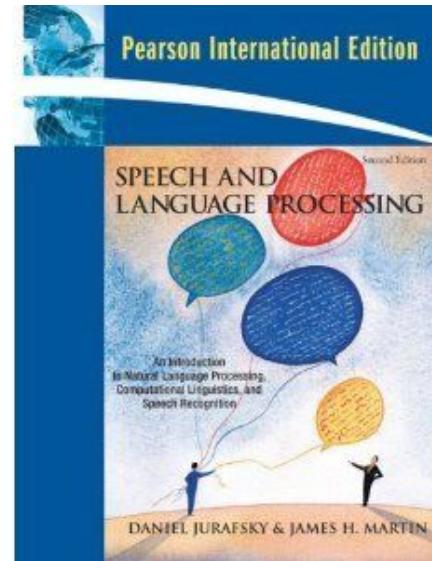
Word Model
Sequence

Tree
Application
Implementation

Date	#	Description
Monday, January 07, 2019	1	Intro; Traditional Tokenization (21 first slides)
Monday, January 14, 2019		Sport Week
Monday, January 21, 2019	2	Tokenization
Monday, January 28, 2019	3	PoS Tagging
Monday, February 04, 2019	4	Language Model
Monday, February 11, 2019	5	Word Representation
Monday, February 18, 2019	6	Text Categorization
Monday, February 25, 2019	7	Midterm Exam
Monday, March 04, 2019		Midterm Exam Week (4-8 Mar)
Monday, March 11, 2019	8	Parsing
Monday, March 18, 2019	9	Attention mechanism & Machine Translation (MT1)
Monday, March 25, 2019	10	MT2 (Transformer) & QA
Monday, April 01, 2019	11	Recent Research in NLP
Monday, April 08, 2019	12	Industry Talk
Monday, April 15, 2019		Songkran Festival
Monday, April 22, 2019	13	NLP Application/Paper Presentation & Progress Report
Monday, April 29, 2019	14	Project Presentation
Monday, May 06, 2019		Final Exam Week (6-21 May)

Course Grading

- Assignments 35%
- Midterm 35%
- Project 30%



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