CHULA ENGINEERING
Foundation toward Innovation
COMPUTER

Introduction to Natural Language Processing

2110572: NLP SYS

Peerapon Vateekul & Ekapol Chuangsuwanich

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?

Why NLP is difficult?

NLP & Text mining

Deep Learning

NLP Tools

Course Logistics

Google Cloud Demo

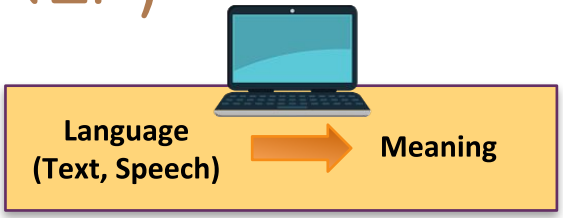
What is NLP?

Definition

Levels of understanding in NLP

NLP today

Natural Language Processing (NLP)



NLP is a subfield in AI, where the goal is

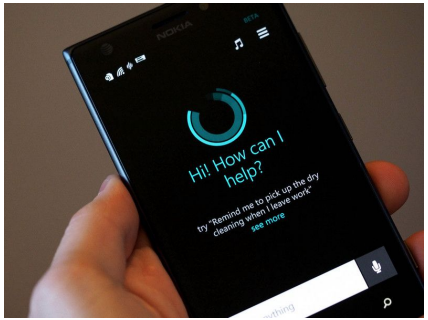
- To bridge the gap between how people **communicate** and what machines **understand** in order to perform useful tasks, e.g. making appointments, buying things, question answering, etc.

NLP (Interpretation)

I want to eat Japanese food.

May I order Yayoi for you?

AI/ML (Generation)

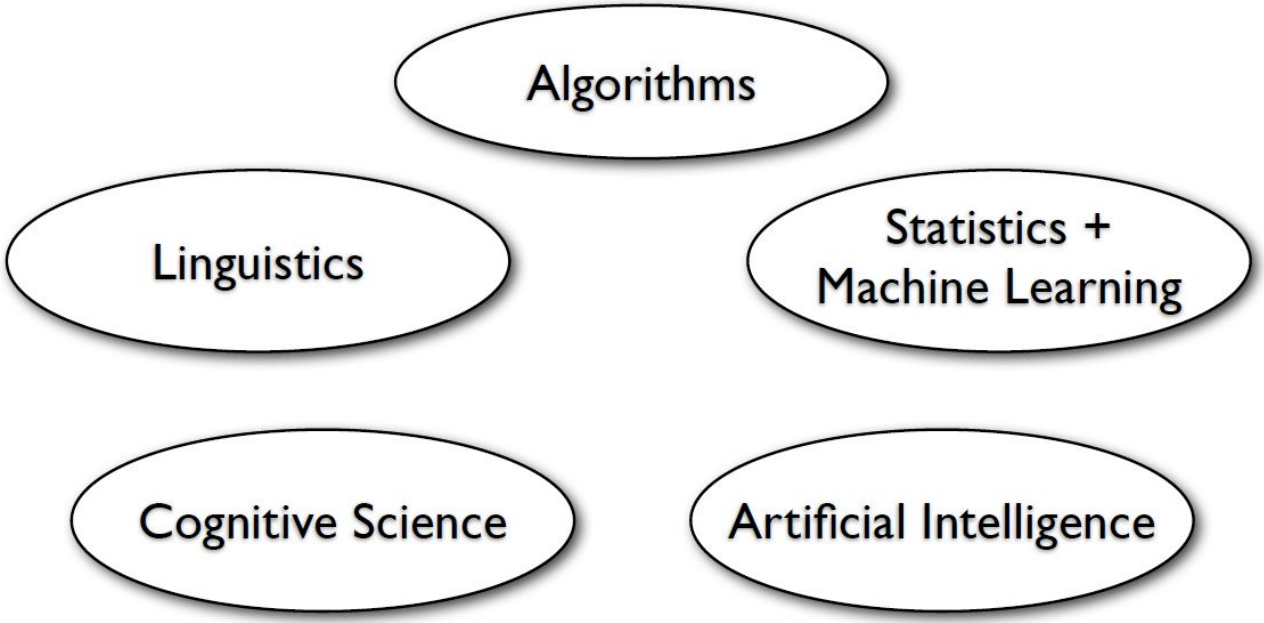


NLP Goals

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

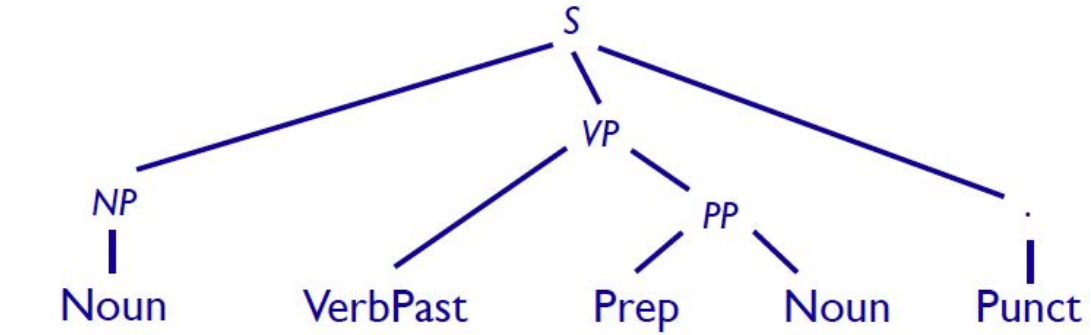
Syntax: Part of Speech

Words

Morphology

Characters

CommunicationEvent(e)
Agent(e, Alice)
Recipient(e, Bob)
SpeakerContext(s)
TemporalBefore(e, s)

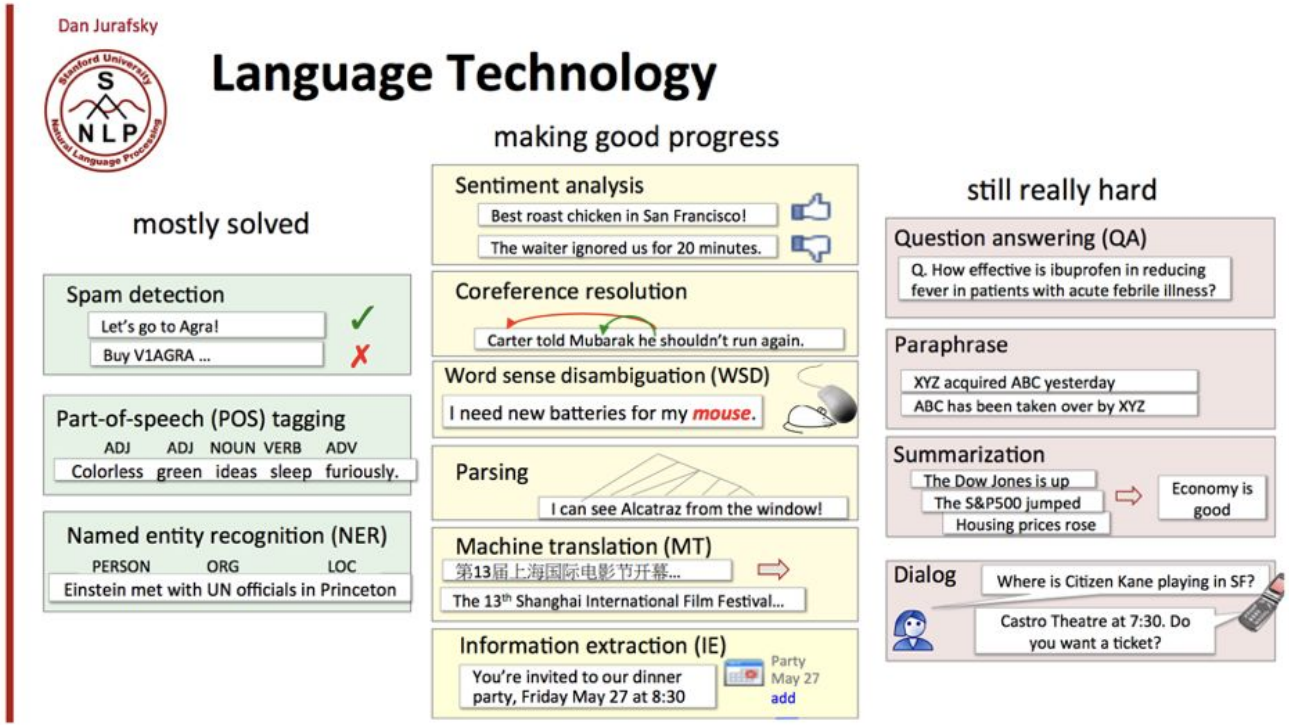


Alice talked to Bob .

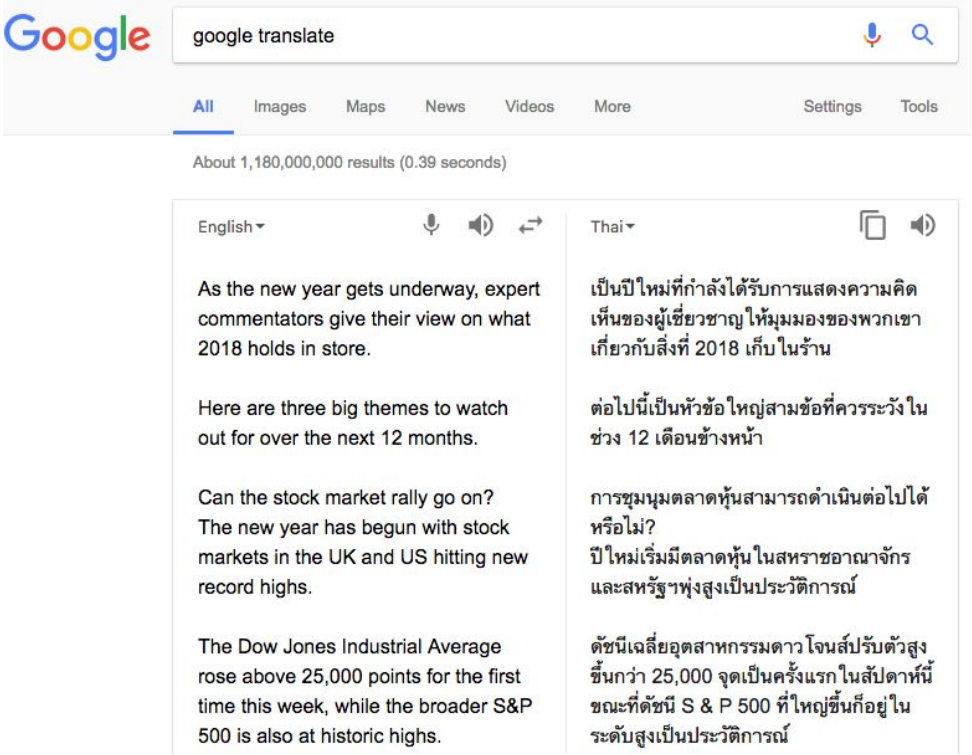
talk -ed [VerbPast]

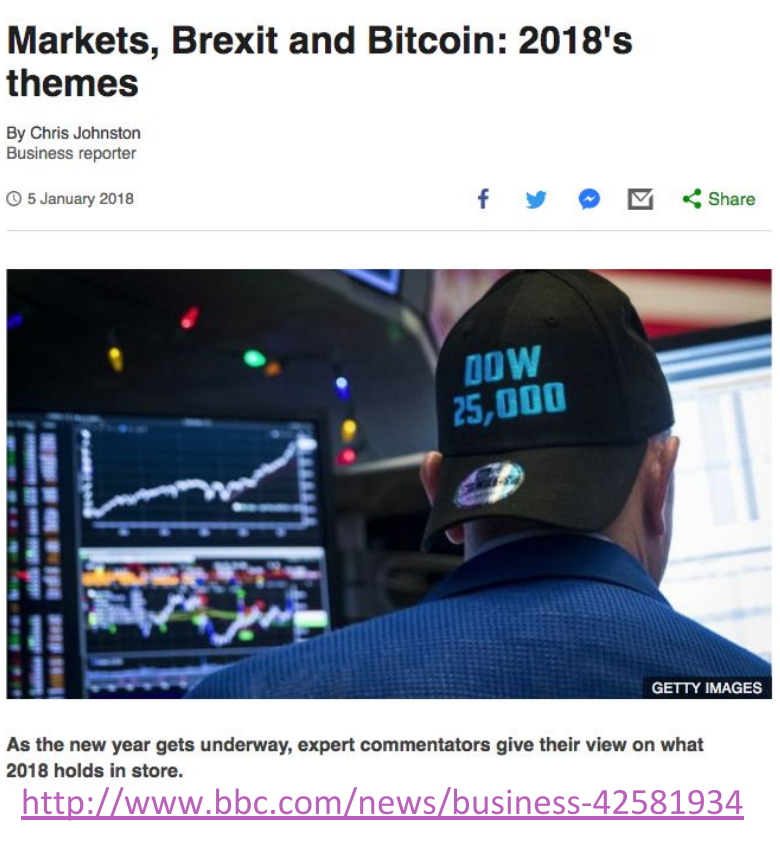
Alice talked to Bob.

NLP today: Technology



NLP today: Machine Translation (MT)





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.

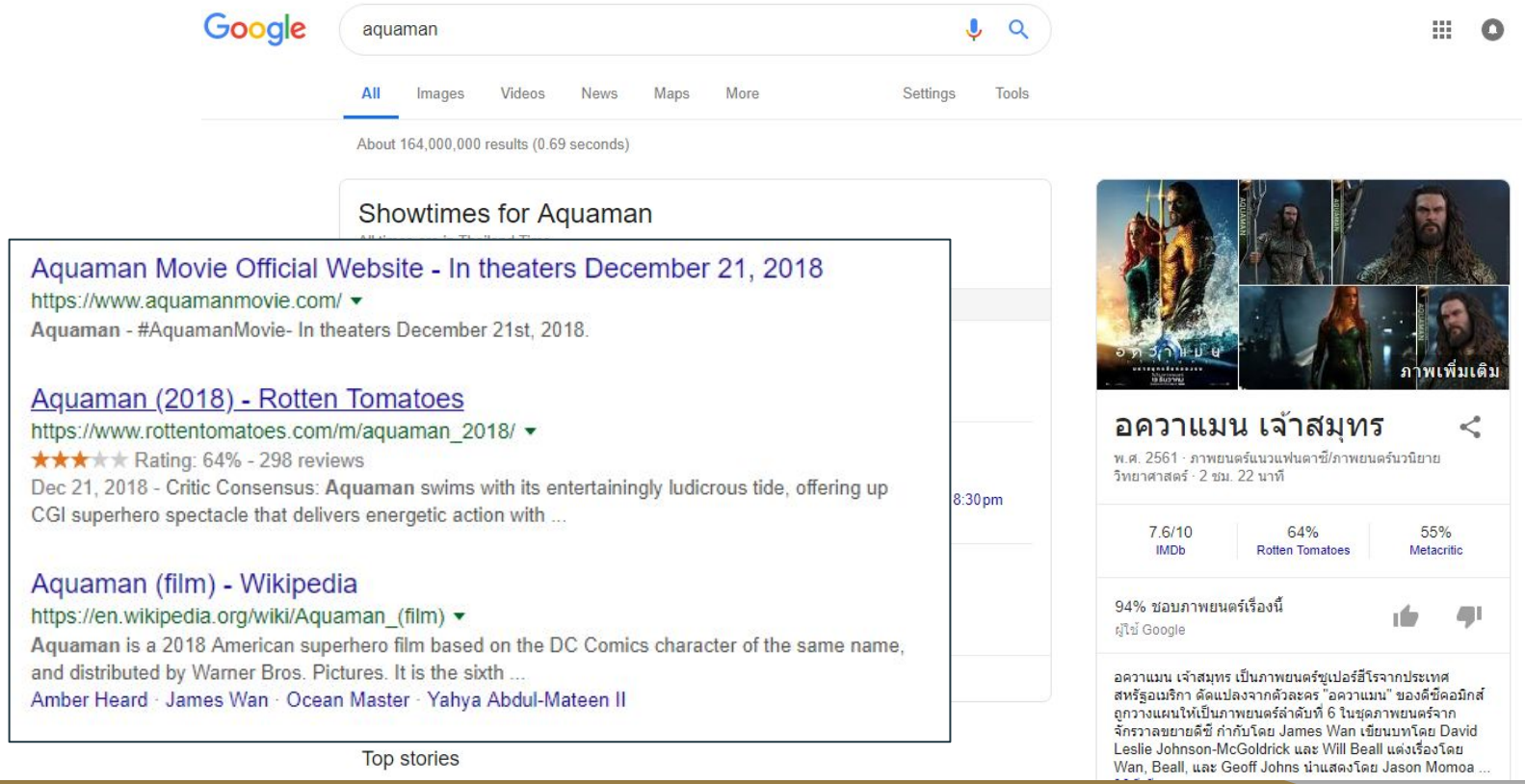
https://www.techrepublic.com/article/ibm-watson-the-inside-story-of-how-the-jeopardy-winning-supercomputer-was-born-and-what-it-wants-to-do-next/

Ref: Prof. Regina Barzilay, NLP @MIT

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

ปริศนาฟ้าแลบ | ตั๊ก, พายไก่, โรเบิร์ต, เป๊ิล | 29 พ.ย. 60 Full HD https://www.youtube.com/watch?v=CAJAQUao7HU

NLP today: Search/Summarization



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	BIDXU-L	0		P	3	CALCS	N
32875	56	1	7/11/06	BIDXB-B	0		N	3		
2247	72	1	4/12/06	BIDXU-R	0		N	3		
45521	61	1	3/30/06	BIDXB-B	0		B	3	CALCS	S
48987	41	1	4/5/06	BIDXB-B	0		P	3	CALCS	N
4179	67	1	5/12/06	BIDXB-B	0		P	2	CALCS	N
26300	59	1	3/31/06	BIDXU-L	0		N	3		
67960	64	1	4/7/06	BIDXU-R	0		P	3	MASS	O
43283	61	W	7/21/06	BIDXB-B	0		B	3		
43219	51	1	4/7/06	BIDXB-B	0		N	3		

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

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

Parsing pathology reports into database

NLP today: IE
Managing customer data

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

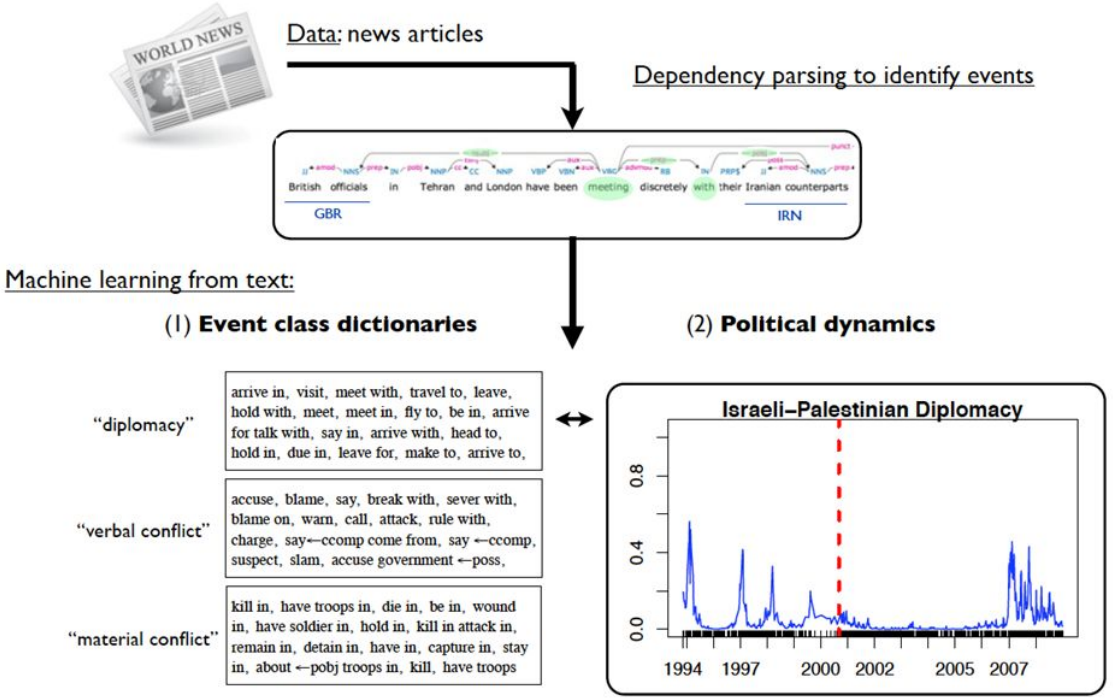
Hugh Son | Tuesday 28 February 2017 11:51 | 16 comments

Like Click to follow The Independent

JPMORGAN CHASE & CO.

- 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.

NLP today: Trend analysis



Hathaway Phenomenon

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

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!

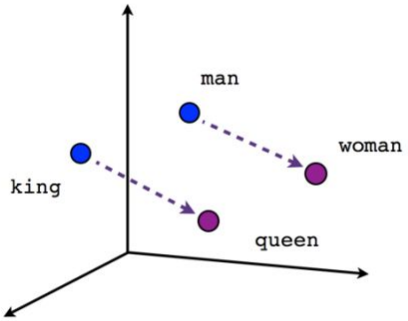
Why NLP is difficult?

Ambiguity

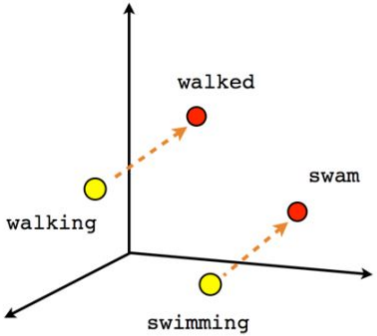
Issues in Thai NLP

What NLP is difficult?

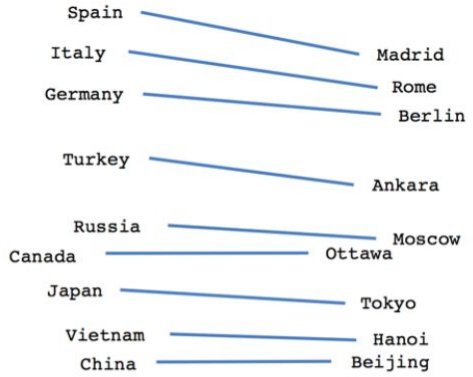
- Complexity in representing, learning and using linguistic/situational/world/visual knowledge



Male-Female



Verb tense



Country-Capital

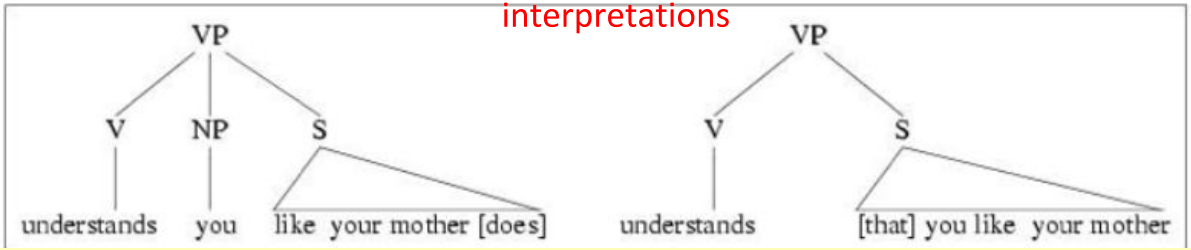
Ref: Prof. Christopher Manning, CS224N/Ling284, 2017

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 interpretations



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

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

Issues in Thai NLP

Word segmentation

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

Sentence segmentation

- No sentence boundary markers

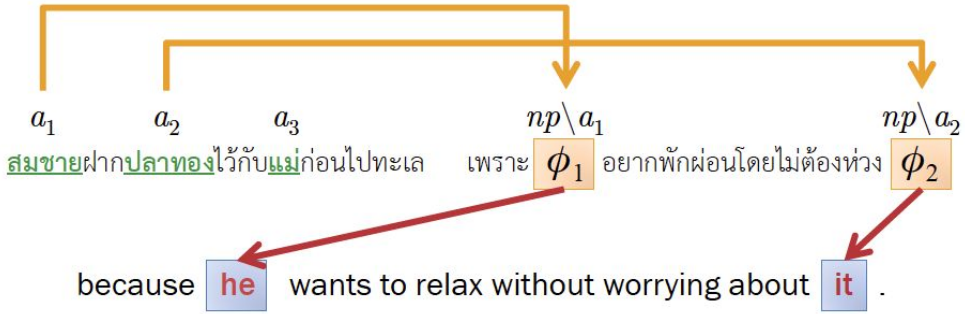
อย่างไรก็ตาม อดีตประธาน ทปอ. กล่าวว่า มีการทักท้วงเรื่องนี้มาตลอดว่า มีช่วงเวลาว่างนานขนาดนี้ ทำไมถึงยังต้องมีการจัดสอบนอกเหนือจากนี้อีก เพราะการสอบล่วงล้ำไปในเวลาระหว่างเรียนมัธยมนั้นกระทบกับเรื่องอื่นๆ โดยเฉพาะการเรียนในชั้นเป็นวงจรลูกโซ่ แนวโน้มที่เข้ามาแก้เรื่องนี้ เป็นความคิดที่ดี แต่ยังไม่เห็นเรื่องใช้ผลการเรียนในชั้นมาเป็นองค์ประกอบรับตรง ซึ่งอาจทำให้เด็กไม่สนใจห้องเรียน และมุ่งกวดวิชา ทำให้การสอบเข้าอุดมศึกษาตกเป็นจำเลยข้อหาทำลายระบบการศึกษาขั้นพื้นฐาน วนไปสู่ปัญหาเก่าๆ ได้

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

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

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

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

Parse → Stop → Stem

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

NER

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

Named Entity Tags

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

Application

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

Tokenization

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

Part of Speech tagging

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

PENN Part Of Speech Tags

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

Ref: BEST2010 dataset

NER

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

Named Entity Tags

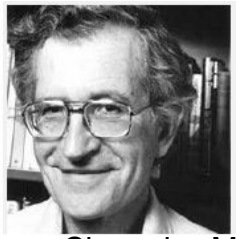
- PER –Person
- LOC – Location
- ORG – Organization
- O – Other

Application

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

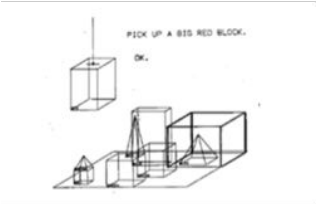
History of NLP techniques

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. (does it)
Person: GRASP THE PYRAMID.
Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
Person: FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX.
Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.
Computer: OK. (does it)
Person: WHAT DOES THE BOX CONTAIN?
Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.
Person: WHAT IS THE PYRAMID SUPPORTED BY?
Computer: THE BOX.
Person: HOW MANY BLOCKS ARE NOT IN THE BOX?
```

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

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

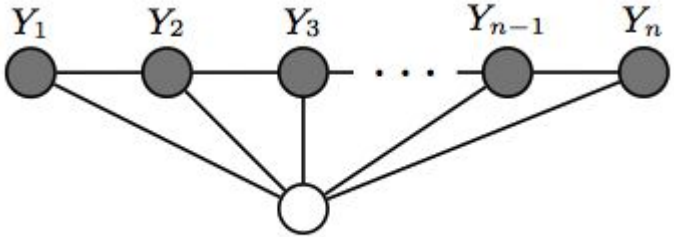
The Internals of SHRDLU

Ref: Prof. Regina Barzilay, NLP @MIT

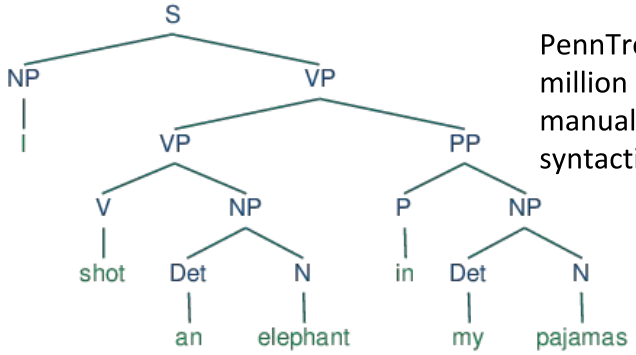
History of NLP techniques

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)



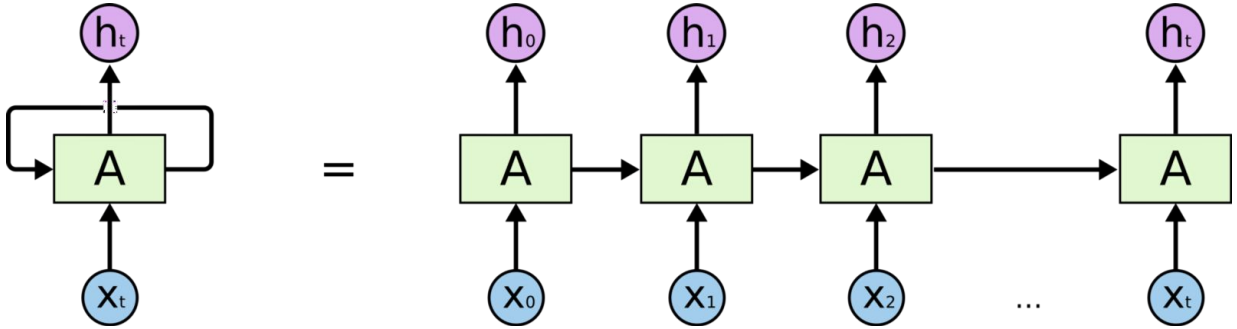
PennTree Bank (1993): one million words from WSJ, manually annotated with syntactic structure

Ref: Prof. Regina Barzilay, NLP @MIT

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.



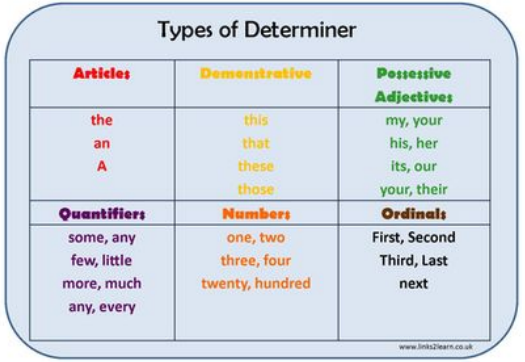
LSTM

Ref: Prof. Regina Barzilay, NLP @MIT

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.



Ref: Prof. Regina Barzilay, NLP @MIT

Case Study: Determiner placement (cont.)

1) Symbolic approach

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

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

- Hard to manually encode this information!

Ref: Prof. Regina Barzilay, NLP @MIT

Case Study: Determiner placement (cont.)

2) Statistical approach

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

"lazy monkeys"

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

-1

"the United States"

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

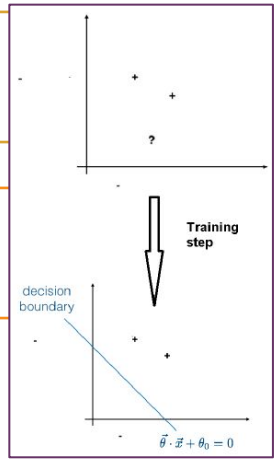
+1

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

Ref: Prof. Regina Barzilay, NLP @MIT

Limitation of traditional statistical approach

1. • Data collection
   • Obtain the labels {-1, +1}
2. • Feature Extraction
   • Most common features in NLP is word n-grams → encoded as one-hot vector
3. • Training
4. • Testing



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

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

pear $[1\ 0\ 0\ 0\ \ldots\ 0]$ → $[0.4\ 0.1\ 0.1]$

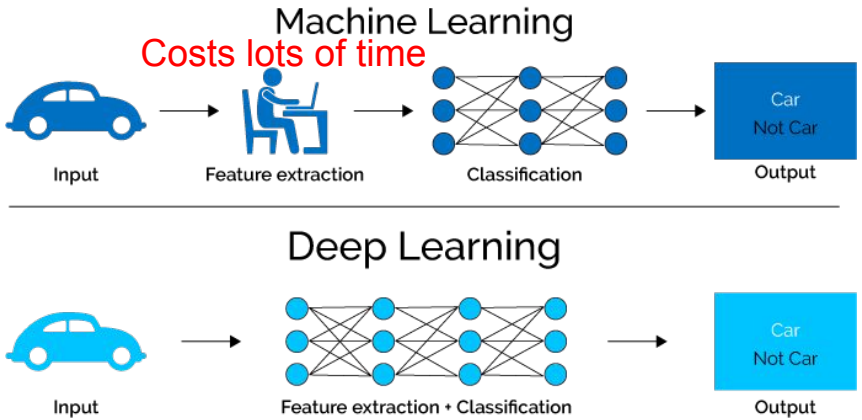
apple $[0\ 0\ 1\ 0\ \ldots\ 0]$ → $[0.6\ 0.2\ 0.3]$

Deep Learning

What is Deep Learning?

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

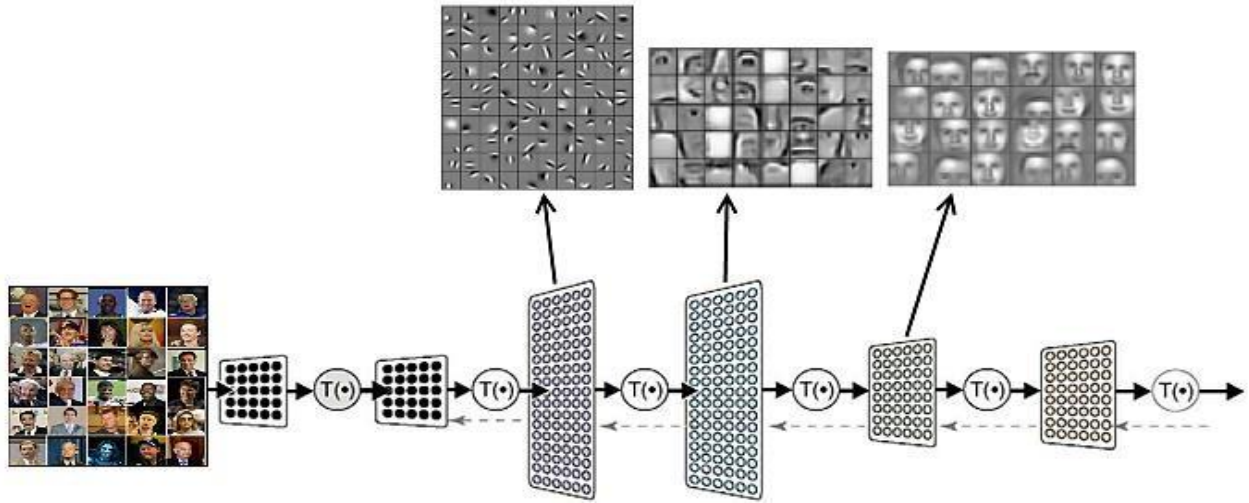
Utilizes learning algorithms that derive meaning out of data by using a hierarchy of multiple layers that mimic the neural networks of our brain.



Deep Learning – Basics (cont.)

What does it learn?

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

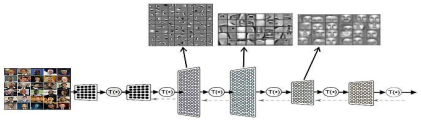


Deep Learning Application

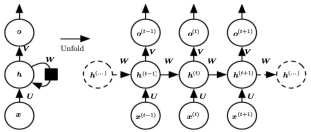
Speech Recognition

Computer Vision

Natural Language Processing



CNN



RNN (LSTM)

NLP + Deep Learning = Deep NLP

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

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]

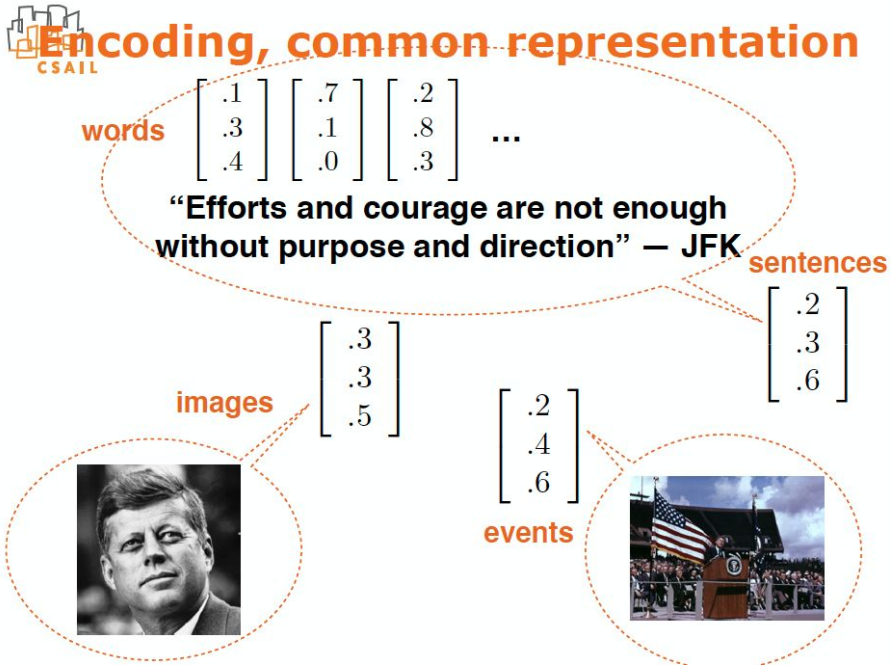
AttaCut: A Fast and Accurate Neural Thai Word Segmenter

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 Segmenter 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 segmenter, which yields syllable embeddings to be used as features by the word segmenter.

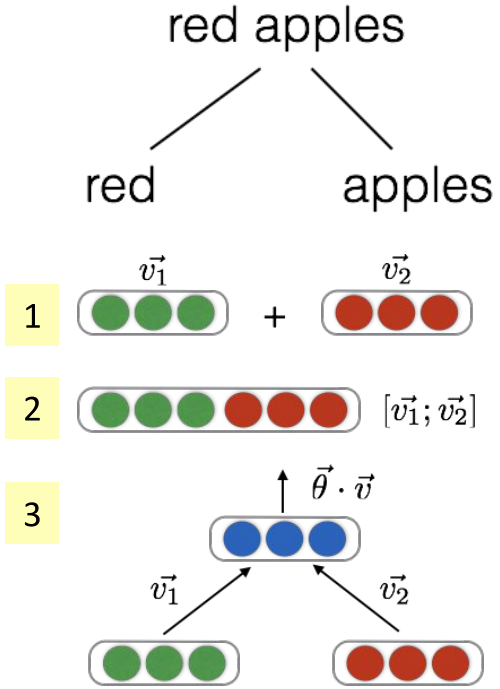
Reasons for exploring Deep Learning

- 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

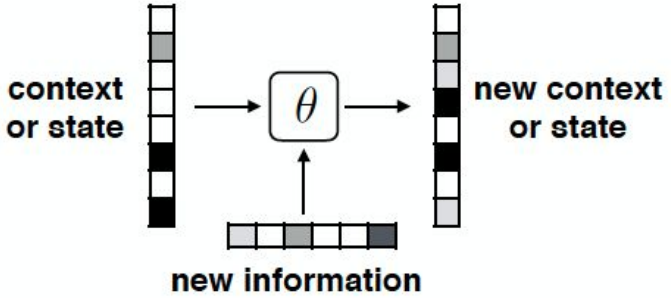


Reasons for exploring Deep Learning (cont.)

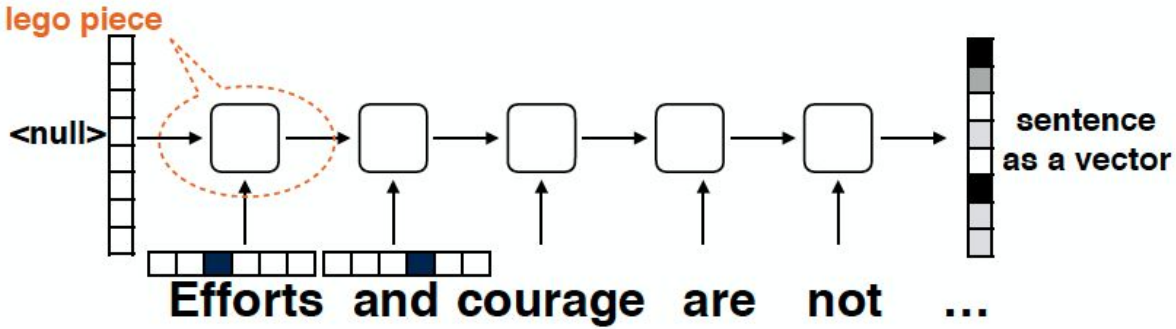
- Flexible neural "Lego pieces"
  - Common representation, diversity of architectural choices
- Can represent any levels of NLP
  - Word
  - Phrase
  - Sentence
  - Paragraph (document)



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

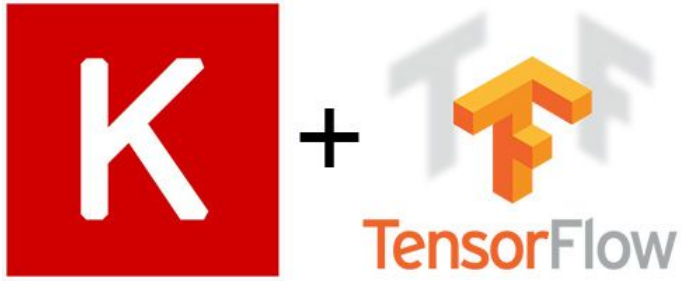


RNN
(LSTM, GRU)



NLP Tools

Implementation



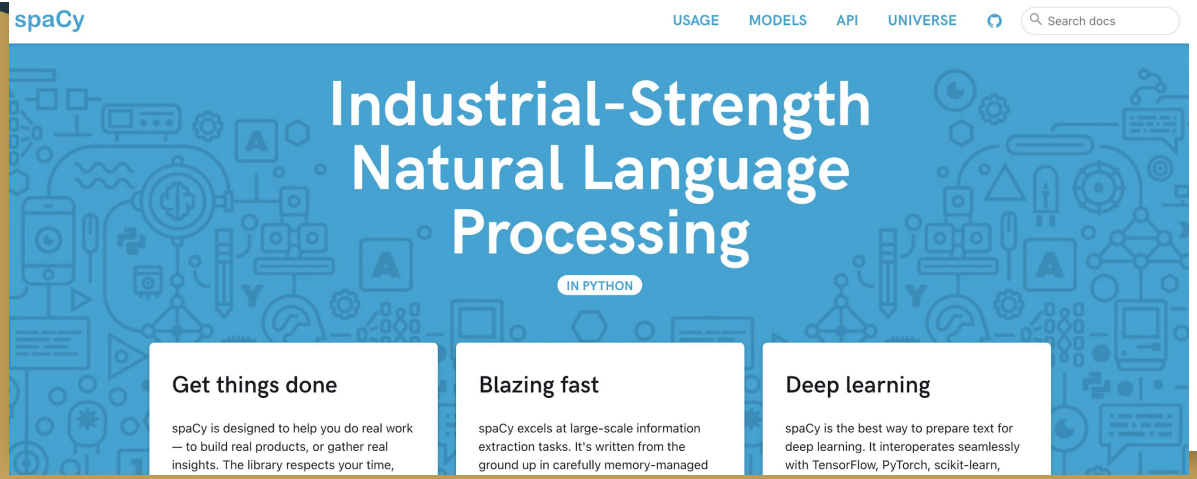
NLP Libraries

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.



HUGGING FACE

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

Star 39,335

Search model

t: Most downloads

- text-classification
- token-classification
- question-answering
- multiple-choice
- masked-lm
- causal-lm
- summarization
- translation
- conversational
- zero-shot-classification
- ExBERT

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

NLP Libraries for Thai

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.

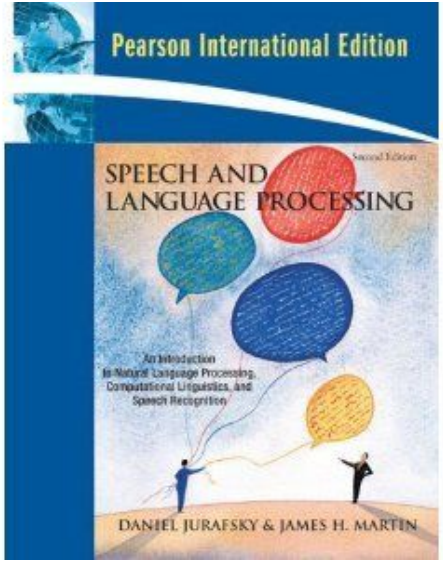
Course Logistics

Class Schedule

Mon (9:30-12:30)	Sat (9:00-12:00)	#	Description
18-Jan-2021	23-Jan-2021	1	Intro; Traditinal 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 + ประกาศ midterm exam (take-home)
1-Mar-2021	6-Mar-2021	7	Parsing
8-Mar-2021	13-Mar-2021		Midterm Exam Week (8-12 Mar)
15-Mar-2021	20-Mar-2021	8	Midterm Exam Submission Attention mechanism & Machine Translation + QA
22-Mar-2021	27-Mar-2021	9	Transformer
29-Mar-2021	3-Apr-2021	10	Recent Research in NLP 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 data: 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: