“NLP”

NLP คือ ความพยายามที่จะแทนที่คน ในการเขียน อ่าน พูดคุย ทำความเข้าใจ ทุกอย่างที่เกี่ยวกับภาษา
Natural Language processing is a technique for automating linguistic tasks.

The diffirante between
Computational Linguistics and Natural language processing

ความแตกต่างระหว่าง CL และ NLP อยู่ที่ใน CL มุ่งเน้นการศึกษาและเข้าใจโครงสร้างของภาษา ในขณะที่ NLP มุ่งเน้นการสร้างโมเดลและเครื่องมือที่ใช้ในการประมวลผลและทำงานกับภาษาธรรมชาติโดยใช้เทคโนโลยีคอมพิวเตอร์ต่าง ๆ และวิธีการทางคณิตศาสตร์ในการทำงานกับภาษาธรรมชาติในลักษณะที่เครื่องคอมพิวเตอร์สามารถเข้าใจและประมวลผลได้

summary
Natural Language Processing

- **เทคนิควิธีสำหรับการใช้เครื่องคอมพิวเตอร์ประมวลผล วิเคราะห์ ข้อมูลภาษา จำนวนมาก**
- **เทคนิควิธีสำหรับการ automate หน้าที่ทางภาษาต่าง ๆ**

Linguistics + Programming + Computer Science = Natural Language Processing !

Applications of Natural Language Processing
1. Text Analytics - discovering insights from text data
2. Artificial Intelligence products that help automate language tasks.

เนื้อหาทั้งหมดอ้างอิงมาจาก ภาษาศาสตร์คอมพิวเตอร์ Thai NLP จุฬาลงกรณ์มหาวิทยาลัย ปี 2564 ของ ผศ. ดร.อรรถพล ธำรงรัตนฤทธิ์

Module 1 Text Classification

Sentiment Analysis การวิเคราะห์ทัศนคติ
ex.

- Market research คนชอบอะไรกันอยู่

** mandala social listening

Detect Rule Comment

- fillter comment

<u>Theme in this course</u>

Machine Learning เป็นแกนสำคัญในการทำ NLP โดยให้เครื่องเรียนรู้ Language task ต่างๆ จากข้อมูลโดยตรง

- การวิเคราะห์ข้อมูลภาษาโดยใช้หลักการทางภาษาศาสตร์ทำให้พัฒนาระบบให้แม่นยำขึ้นได้

Module 2: Language Modeling

- ex. Spell and Grammar Checker
- predictive keyboard

- Chat GPT !

Module 3 : Search

- Information Retrieval

เทคโนโลยีการค้นคืนข้อมูลด้วยคอมพิวเตอร์ทำให้คนเข้าถึงข้อมูลได้ง่ายขึ้น

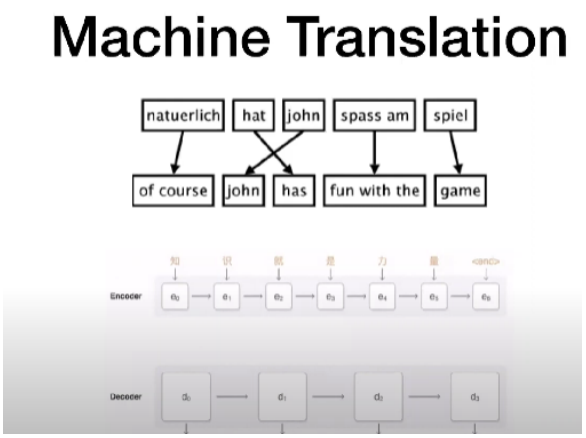
Module 4: Information Extraction

- คือการทำ Text เข้าเข้าใจมากขึ้น
- ใคร ทำอะไร ที่ไหน

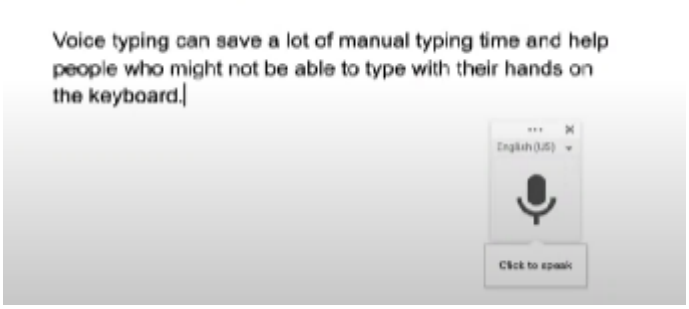
ยาก เพราะ ต้องอิงถึงโลกแห่งความเป็นจริง เช่น ลุงตู่คือ นายก จริงๆ เมื่อเช้าหมายถึงเมื่อไร

Module 5: Language Generation

- Machine Translation



- Speech Recognition



Chat Gpt Again Haha

NPL Systems
ex
โปรแกรมสอนภาษา
โปรแกรม Chat Bot
โปรแกรมจับ Fale news
โปรแกรมตรวจจับความยากของภาษา
โปรแกรมวิเคราะห์ Resume เพื่อหางานที่เหมาะสม

Text Processing

what is text ? รูปแบบของ String การเข้ารหัส Unicode
รวมถึงการร้อยเรียง emoji เป็นประโยค

Issues with text data

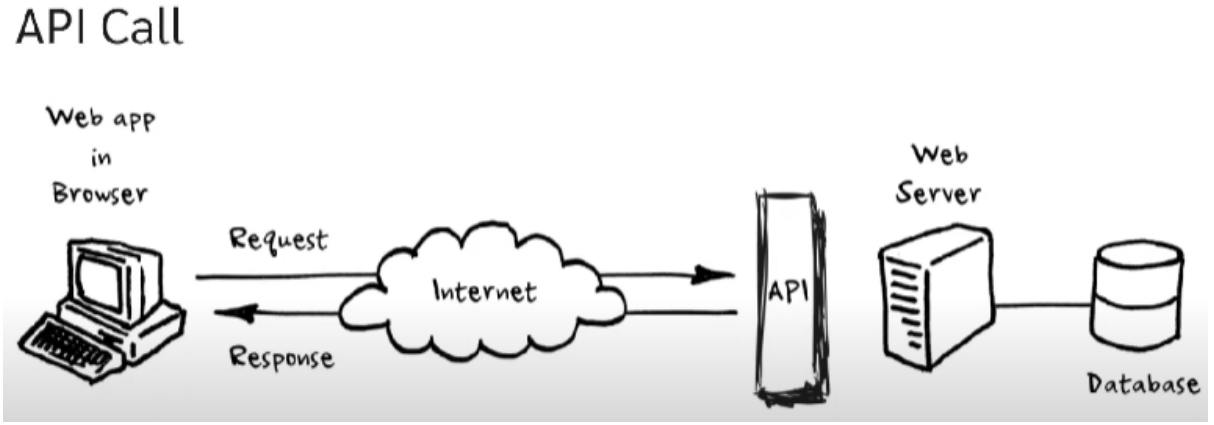
1. Noise from the date collection chanel
2. Variation of languages (Sociolinguistic factors)
3. Complexity of language

Noise ??

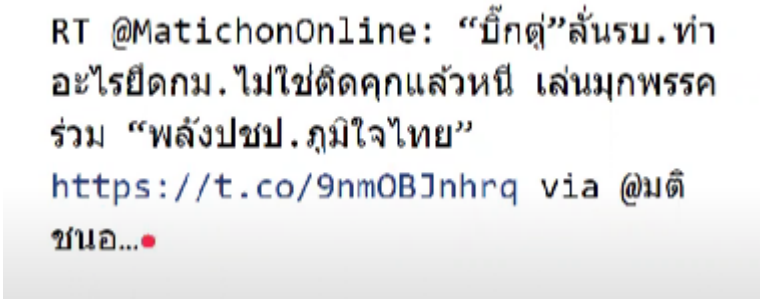
Where do text come from ?

- A file
  mistake form the author
  human error

- The internet
  noise from internet less than a File

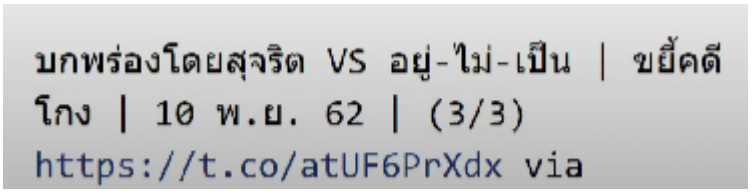


but still have
such a



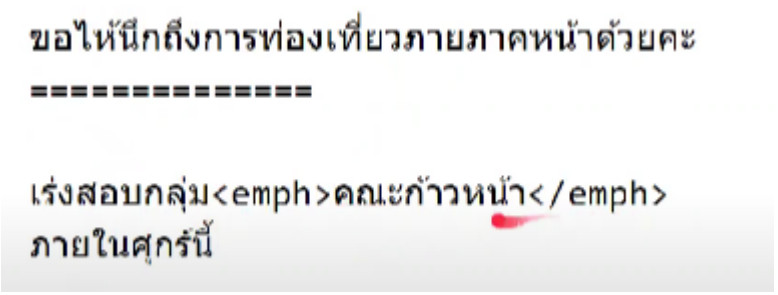
what is ร.บ ก.ม

and



the data is messy !!!

and



What we do

Clan out the non-text portion such as URL, #, numbers, HTML tags

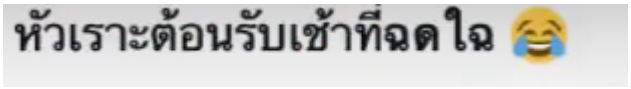
Decide whether you want other languages too
หากติดภาษาอื่นเข้ามา จะเอาไหม

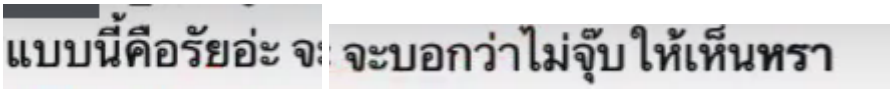
Check the vocabulary to see if there is any remnants

Next

Variation of languages (sociolinguistic factors) * ภาษาไทยดิ้นได้

การแปรง
ex.
Social media such as Twitter , Factbook , and instagram is data **stream** which provides cheap and up to date opinions of the public.
ดึงออกมาง่าย และ สดใหม่
informal





Complexity of language

Language has structures, To analyze language data, we must analyze its structure
ภาษามีโครงสร้าง เพื่อวิเคราะห์ข้อมูลภาษาเราต้อง วิเคราะห์โครงสร้าง

- Discourse – a list of sentences ตัดประโยค วิเคราะห์เป็นทีละก้อนๆ
- Syntax – sentence and phrase structure
- Morphology – structure of a word

Common linguistic analysis การวิเคราะห์ทางภาษาทั่วไป

In NLP we call this process linguistic preprocessing

- **Word segmentation**
- **Sentence segmentation**
- **Morphological analysis/Normalization**

<u>Word Segmentation</u>

- •Dictionary-based system with **maximal matching** algorithm e.g.
- •the ‘newmm’ engine on pythainlp

- •Machine learning system e.g. ‘deepcut’ ‘oskut’ attacut’ on pythainlp

Maximal Matching Algorithm

- •**you need a good dictionary:** good coverage but not too broad . what does this mean?
  ครอบครุม แต่ต้องไม่กว้างเกิน เช่น คำโบราณที่อาจไม่ได้ใช้

•Generate all possible segmentations and select the combination with the fewest word (each word must be maximally long)
ตัดให้ครบทุกแบบ และหาคำที่ตัดน้อยที่สุด เช่น

แต่มีคำอีกเยอะที่ผิด เช่น กอดอก จะออกมาเป็น กอด/อก และ กอ/ดอก ตัด2เท่ากัน

Machine Learning system

•Machine learning is a family of algorithms that learn certain task form data

•Segment a large amount of text by hand → training sample
Use a machine learning model to learn from training samples

ให้ตัวอย่างการตัดคำที่ถูกต้อง และ ให้ เครื่องเรียนรู้จากข้อมูลที่ฝึกเข้ามา

Method	Dataset (WL_{F_1})		
	In-Domain	Out-Domain	
	BEST-2010	Wisesight	TNHC
Previous work			
Dictionary-based	71.18%	78.97%	72.70%
DeepCut	94.46%	84.45%	78.17%
Ours			
BiLSTM			
(CH)-BI	95.05%	85.85%	79.31%
(CH+SY)-BI	**95.59%**	**86.15%**	78.70%
-CRF(SY)-BI	95.51%	86.10%	**79.89%**
ID-CNN			
(CH)-BI	94.31%	85.80%	79.22%
(CH+SY)-BI	95.45%	**86.43%**	**79.87%**
-CRF(SY)-SchA*	**95.60%**	86.15%	79.64%

Word Segmentation is very easy

```
import pythainlp, attacut
test_sentence = "ระฆังดีไม่ดีก็ดัง"

tokens = pythainlp.word_tokenize(test_sentence, engine="attacut")
```

ถ้าไม่ใส่ engine จะเป็น Dictinaly

```
5] pythainlp.word_tokenize('พระมหาไพรวัลย์ถึงกับต้องบอกให้เจเจยิ้มแบบอรุ่มเจ๊าะอีกรอบ', engine='attacut')

['พระมหาไพรวัลย์',
 'ถึง',
 'กับ',
 'ต้อง',
 'บอก',
 'ให้',
 'เจเจยิ้ม',
 'แบบ',
 'อรุ่มเจ๊าะ',
 'อีก',
 'รอบ']
```

```
import nltk
nltk.download('punkt')
```

```
doc = nltk.word_tokenize("Apple is looking at buying U.K. startup for $1 billion")
doc
```

```
['Apple',
 'is',
 'looking',
 'at',
 'buying',
 'U.K.',
 'startup',
 'for',
 '$',
 '1',
 'billion']
```

Sentence segmentation ตัดประโยค

•ประโยคจะถูก defind ด้วย Gramma เช่น N + V.1 + กรรม

Sentence segmentation is also known as sentence boundary detection.
It is very important task for NLP because it allows us to process a smaller and meaningful chunk of text (instead of the entire long chink)
หาว่าตัวแบ่งประโยค อยู่ตรงไหน

How do we do sentence segmentation for English?

Regular expression

! ? are very good boundaries , but . is very ambiguous

ex Dr.Attapol doesn't work at Google Inc.

I think…the accuracy is 10.5%

Tokenize and use some complicated set of rules and an abbreviation dictionary.
ใช้งานและใช้ชุดกฎที่ซับซ้อนและพจนานุกรมคำย่อ
ดังนั้นเราจึงต้องสร้างกฎหรือใช้ เทคนิคต่างๆเข้ามาช่วย เช่น

Machine Learning (Kiss and Strunk, 2006)

•Give a large amount of data , find the collocations or words that appear together a lot . Those should not be separated into two sentences.
คำที่ปรากฏพร้อมกันจำนวนมาก สิ่งเหล่านี้ไม่ควรแยกออกเป็นสองประโยค

•Find words that are at the beginning of the text (hence beginning of the sentence) a lot
หาคำที่อยู่หน้าข้อความ (ซึ่งเป็นจุดเริ่มต้นของประโยค) บ่อยๆ

Accuracy rates are very high

It turns out to be effective for many languages that ambiguously mark their sentence boundaries.

Corpus	Error (<S>) (%)	Prec. (<S>) (%)	Recall (<S>) (%)	F (<S>) (%)
B. Port.	1.11	99.14	99.72	99.43
Dutch	0.97	99.25	99.72	99.48
English	1.65	99.13	98.64	98.89
Estonian	2.12	98.58	99.07	98.83
French	1.54	99.31	99.08	99.19
German	0.35	99.69	99.93	99.81
Italian	1.13	99.32	99.49	99.41
Norw.	0.81	99.45	99.68	99.56
Spanish	1.06	99.66	99.23	99.45
Swedish	1.76	98.82	99.36	99.09
Turkish	1.31	99.40	99.24	99.32
Mean	1.26	99.25	99.38	99.31
SD	0.49	0.33	0.38	0.29

Sentence segmenting

```
import nltk
nltk.download('punkt')
from nltk.tokenize import
PunktSentenceTokenizer

text = "this is a sentence. And this is
another one! are you enjoying it?"
tokenizer = PunktSentenceTokenizer()
sentences = tokenizer.tokenize(text)
print(sentences)
```

```
['this is a sentence.', 'And this
is another one!', 'are you enjoying
it?']
```

Morphological analysis/Normalization

เช่น Verb เปลี่ยนได้กี่รูป
seek seeked sought seeking seeker
ต้องหาวิธี map ให้เป็นตัวเดียวกัน
บางคำมี s เติมหลังอยู่แล้วไม่ได้ระบุว่าเป็นคำพหุพจน์