

บทที่ 4 - NLP คืออะไร บทเรียนจากอดีตสู่ปัจจุบัน

Peerat Limkonchotiawat

PhD student at VISTEC, Thailand

สารบัญ

เราจะเรียนอะไรกัน在本นี้

NLP คืออะไร?

ประวัติศาสตร์ NLP

NLP ภาษาไทย



NLP คืออะไร

NLP คืออะไร

- NLP หรือ Natural Language Processing เป็นสาขาย่อยของภาษาศาสตร์ (Linguistics) และ ปัญญาประดิษฐ์ (Artificial Intelligence)
- ศึกษาค้นคว้าเกี่ยวกับการทำให้คอมพิวเตอร์ “เข้าใจและจัดการ” ภาษารธรรมชาติของมนุษย์ได้
- ยกตัวอย่าง
 - จำแนกประเภทประโยค (Sequence Classification)
 - หนังสือนี้สนุกจังเลย → ประโยคเป็น “บวก”
 - หนังสือนี้แย่มาก → ประโยคเป็น “ลบ”
 - สร้างข้อความ-เติมคำในช่องว่าง (Text Generation)
 - กะเพรา__อร่อยมาก → “กะเพราะหมูสับอร่อยมาก”, “กะเพราะไก่อร่อยมาก”

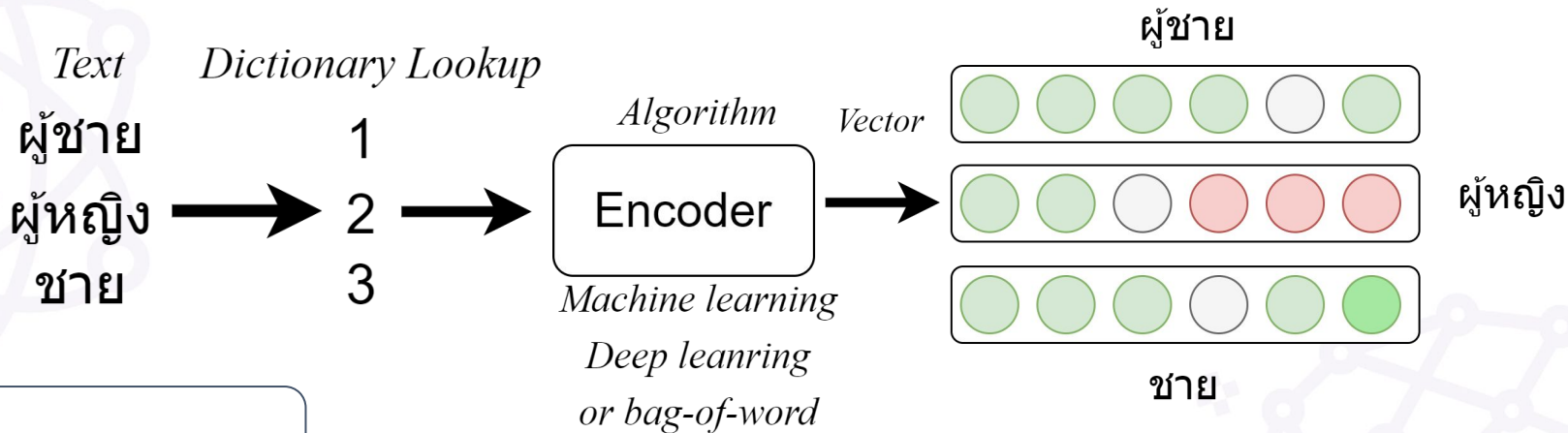
Challenge?

- คอมพิวเตอร์ไม่ได้เข้าใจ “ภาษาแบบที่มนุษย์เข้าใจ”
 - เช่น สองข้อความนี้แตกต่างกันหรือไม่?
 - “ฉันหิวข้าวมากๆ” และ “ฉันอยากรับประทานอาหารมากๆ”
 - มนุษย์: “ไม่”
 - คอมพิวเตอร์: “ใช่”
 - สาเหตุ: “หิวข้าว” != รับประทานอาหาร



คอมพิวเตอร์เข้าใจภาษาคน?

ทำไมคอมพิวเตอร์ถึงเข้าใจภาษาคน?



ฉันไม่เข้าใจตัวอักษร



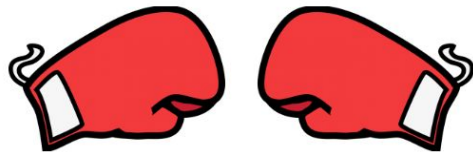
แล้วเลข 1,2,3 มีความ
ยังไง?



ประวัติศาสตร์ NLP (แบบง่ายๆ สั้นๆ ไร้คณิตศาสตร์!!)

ตัดคำ ตัดพยางค์ แล้วทำ Bag-of-Word

- การตัดคำ เป็น task แรกสุดในโปรเซสของการทำ NLP
- หลายภาษาใช้ whitespace ในการแบ่งคำ บางภาษาต้องทำโมเดลขึ้นมาเพื่อตัดคำเช่น ภาษาไทย



ตาก|ลม VS ตา|กลม

ฉันนั่งตาก|ลมอยู่ริมทะเล

บ๊วกายทำให้เธอตา|กลมมาก

ตัดคำ ตัดพยางค์ แล้วทำ Bag-of-Word

หน่วยคำ
(Word)

วันนี้ฉันสั่งกิวดังมาทานที่บ้าน

หน่วยคำย่อย
(Subword)

วันนี้ฉันสั่งกิวดั้ังมาทานที่บ้าน
(SentencePiece; XLMR)

หน่วยพยางค์
(Syllable)

วันนี้ฉันสั่งกิวดังมาทานที่บ้าน

หน่วยตัวอักษร
(Character)

ว้ัน|น|ี|ั้|ฉั|น|สั|ั้|ง|กิ|ว|ด|ั้|ง|มา|ทา|น|ที่|ี|ั้|บ|ั้|า|น

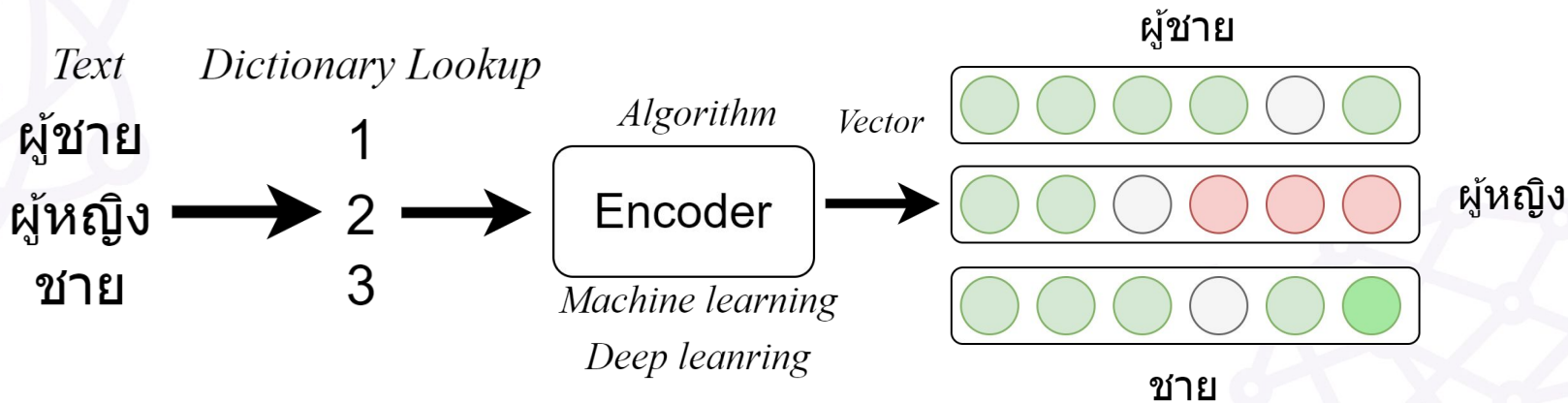
ตัดคำ ตัดพยางค์ แล้วทำ Bag-of-Word

ประโยค	แมว	นั่ง	มอง	ปลา	บน	โต๊ะ	กิน
แมว นั่ง มอง ปลา	1	1	1	1	0	0	0
แมว นั่ง บน โต๊ะ แมว	2	1	0	0	1	1	0
แมว กิน ปลา	1	0	0	1	0	0	1

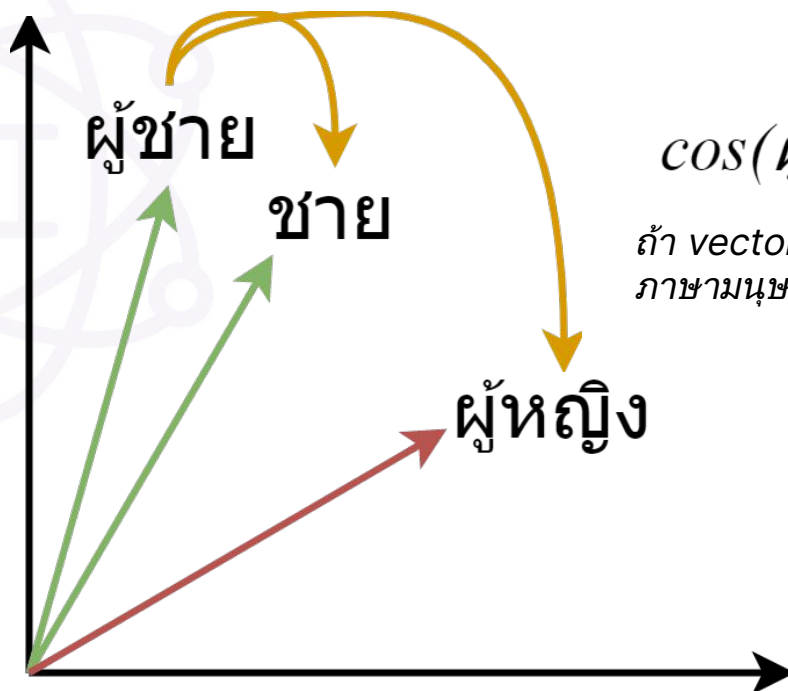
ตัดคำ ตัดพยางค์ แล้วทำ Bag-of-Word

ข้อดี	ข้อเสีย
<ul style="list-style-type: none">• ไม่ยุ่งยาก ทำได้ง่ายสุด• ประสิทธิภาพยังดีมาก แม้ในยุคปัจจุบัน	<ul style="list-style-type: none">• ต้องพึ่งการตัดคำมากไป• BoW ไม่มีลำดับของคำมาเกี่ยว• ไม่รองรับคำที่ไม่เคยเห็น

แปลงข้อความเป็น Vector (Discrete)



แปลงข้อความเป็น Vector



Cosine similarity ($\cos(\cdot)$)

$$\cos(\text{ผู้ชาย}, \text{ชาย}) > \cos(\text{ผู้ชาย}, \text{ผู้หญิง})$$

ถ้า vector ของสามารถแปลงความหมายได้ดี = คอมพิวเตอร์เข้าใจภาษามนุษย์ได้ดี

แปลงข้อความเป็น Vector

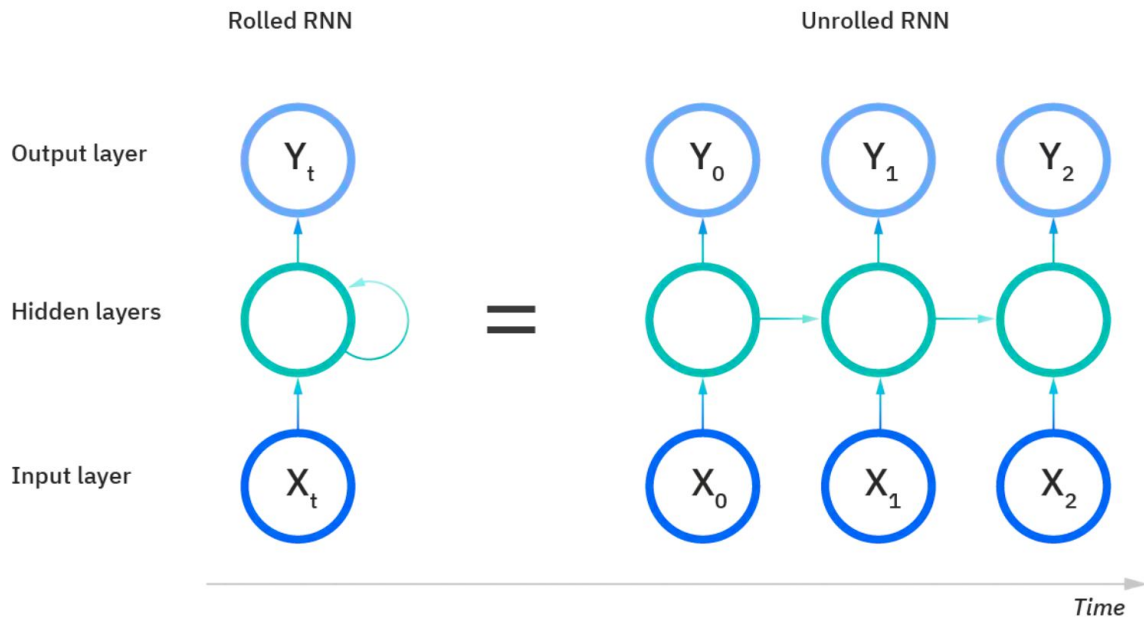
ข้อดี	ข้อเสีย
<ul style="list-style-type: none">มีประสิทธิภาพสูงในงานที่ต้องการความซับซ้อนรองรับคำที่ไม่เคยเห็น	<ul style="list-style-type: none">การเทรนโมเดลมีความซับซ้อนกว่า BoW



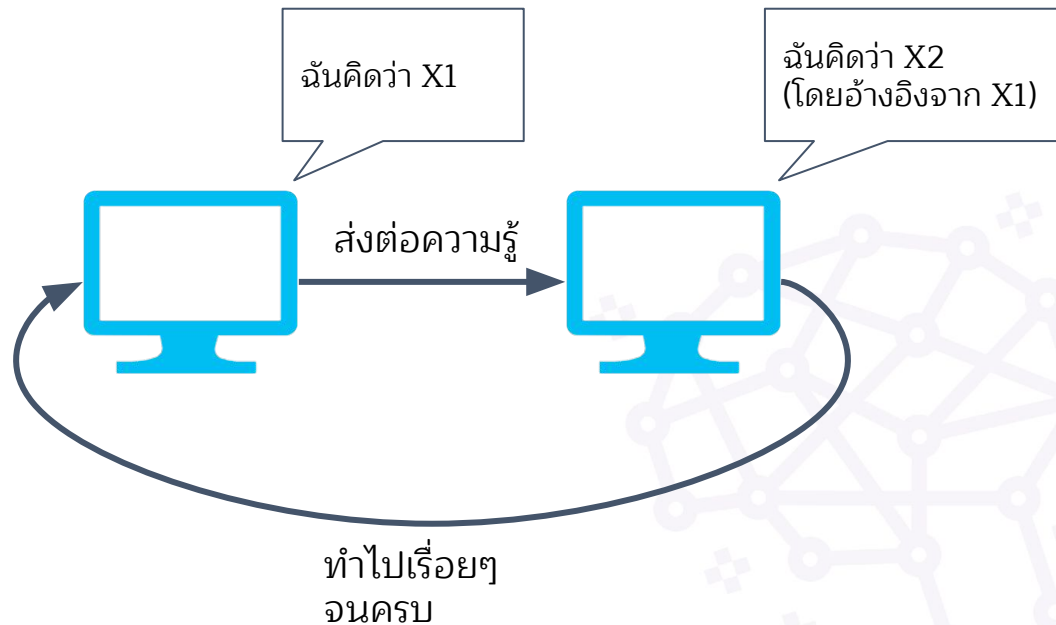
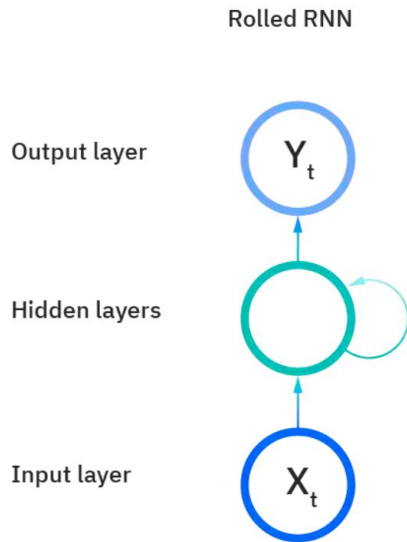
แล้วสร้าง Vector ออกมายังไง?



ยุคทองของ “RNN”



ยุคทองของ “RNN”



ยุคทองของ “RNN”

cstorm125 / thai2fit (Public) Watch 12

<> Code Issues 1 Pull requests Actions Projects Wiki Security

master 1 branch 1 tag

Go to file Add file Code

cstorm125 add citation details 7e58550 on Jan 9, 2021 51 commits

images	add submission image	3 years ago
thwiki_lm	pycon 2019	3 years ago
wongnai_cls	Add USE benchmark	2 years ago
.gitattributes	lfs	4 years ago
.gitignore	wongnai benchmarks all done	3 years ago
LICENSE	Initial commit	4 years ago
README.md	add citation details	14 months ago

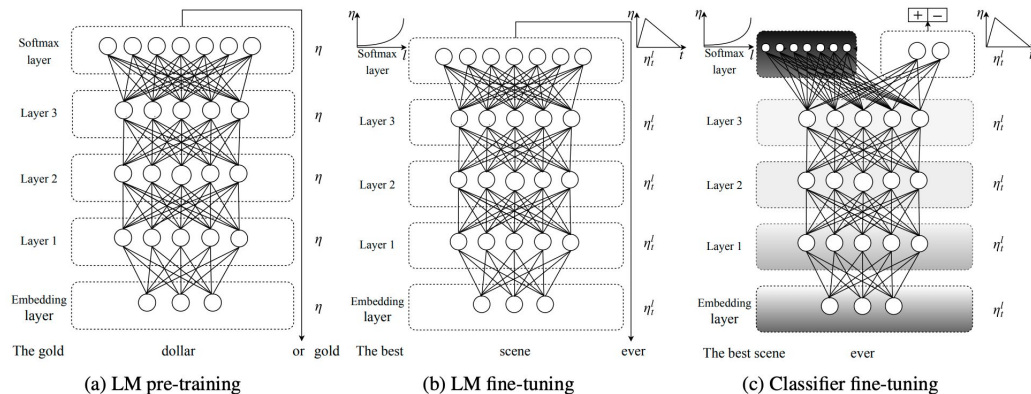
README.md

thai2fit (formerly thai2vec)

ULMFit Language Modeling, Text Feature Extraction and Text Classification in Thai Language.
Created as part of [pyThaiNLP](#) with [ULMFit](#) implementation from [fast.ai](#)

Models and word embeddings can also be downloaded via [Dropbox](#).

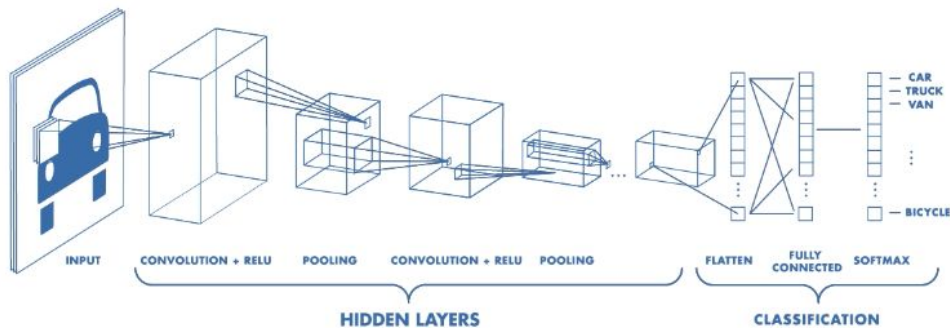
We pretrained a language model with 60,005 embeddings on [Thai Wikipedia Dump](#) (perplexity of 28.71067) and text classification (micro-averaged F-1 score of 0.60322 on 5-label classification problem. Benchmarked to 0.5109 by [fastText](#) and 0.4976 by LinearSVC on [Wongnai Challenge: Review Rating Prediction](#). The language model can also be used to extract text features for other downstream tasks.



ยุคทองของ “RNN”

ข้อดี	ข้อเสีย
<ul style="list-style-type: none">มีประสิทธิภาพสูงในงานที่ต้องการความซับซ้อนลำดับของคำ มีความสำคัญ	<ul style="list-style-type: none">ใช้เวลาในการเทรนโมเดลเป็นเวลานานยิ่งวน ยิ่งลืมของก่อนหน้า หรือยิ่งซับซ้อน

ยุคทองของ “BERT”



ไม่นานเหมือน RNN แต่เหมาะกับ
classification ที่ไม่ใช่ order

Rolled RNN

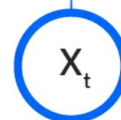
Output layer



Hidden layers



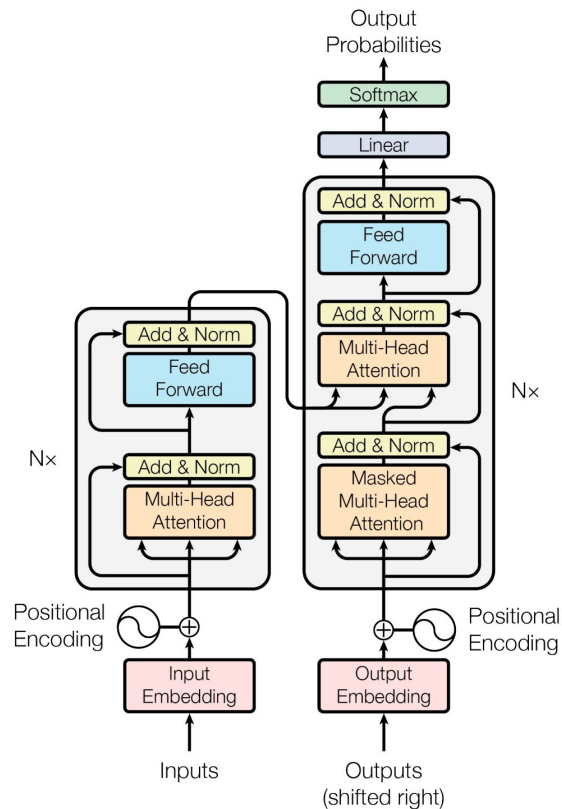
Input layer



เทรนนาน แต่เหมาะกับ classification ที่
order มีผล

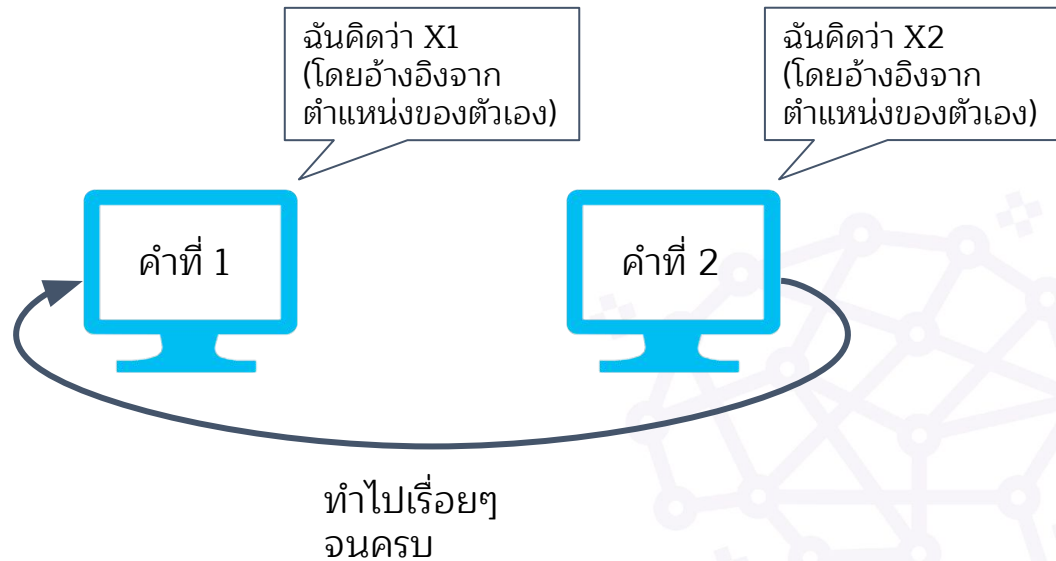
ยุคทองของ “BERT”

- Transformer: รวมข้อดีของ CNN และ RNN
 - คำนวณ output แบบไม่สน order
 - แต่มีการกำหนดตำแหน่งที่ input แทน

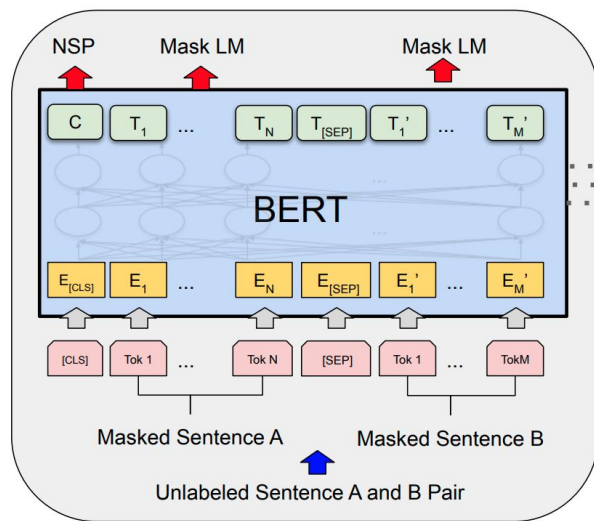


ยุคทองของ “BERT”

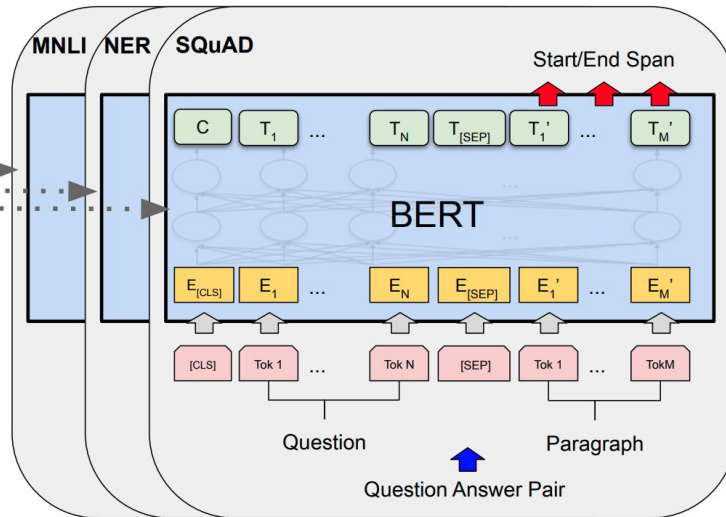
อธิบาย Transformer แบบเข้าใจง่ายๆ



ยุคทองของ “BERT”



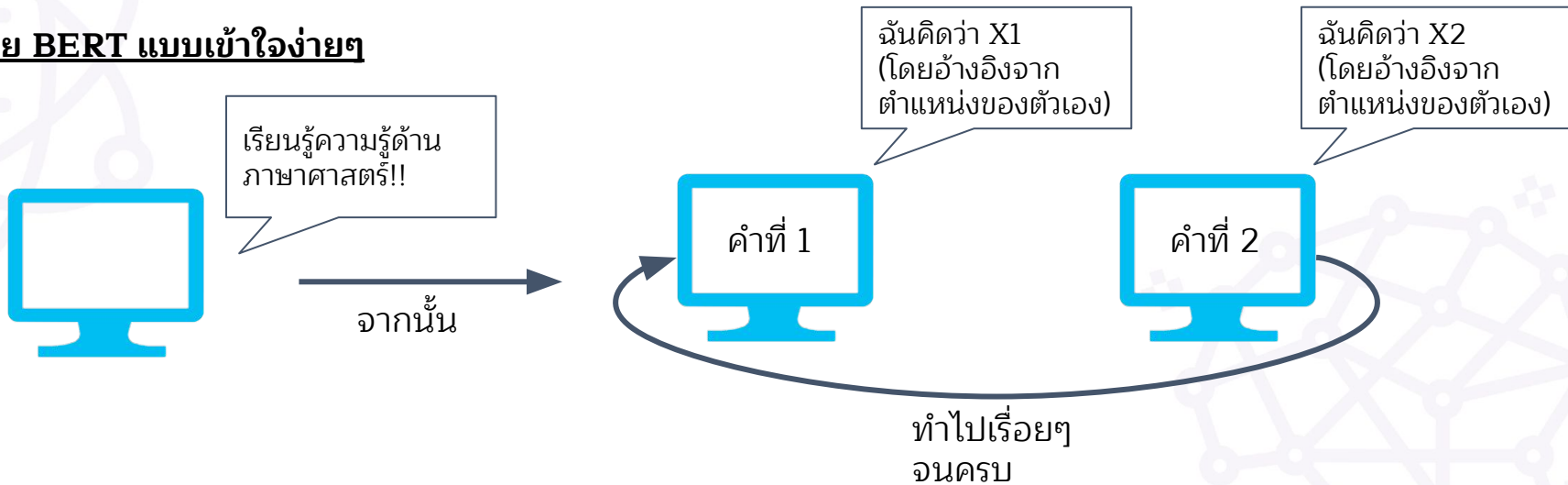
Pre-training



Fine-Tuning

ยุคทองของ “BERT”

อธิบาย BERT แบบเข้าใจง่ายๆ



ยุคทองของ “BERT”



VISTEC-depa AI Research Institute of Thailand

Jan 24, 2021 · 5 min read

WangchanBERTa โมเดลประมวลผลภาษาไทยที่ใหญ่ และก้าวหน้าที่สุดในขณะนี้

เปิดให้ทุกคนใช้ฟรีโดย AIResearch.in.th และ VISTEC ภายใต้สัญญาอนุญาต CC-BY-SA 4.0



Image by Phannisa Nirattiwongsakorn

การจำแนกข้อความภาษาไทย ตั้งแต่ BoW \Rightarrow BERT

Let's code!

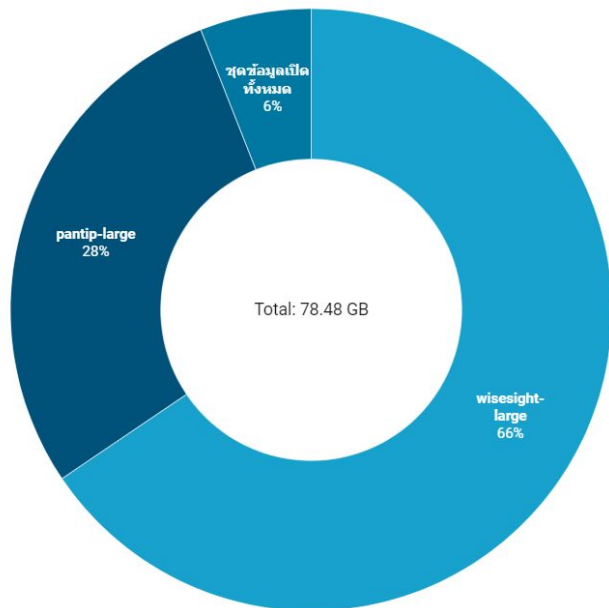


NLP ภาษาไทยไปถึงไหนแล้ว?

โมเดลและคลังข้อมูลสาธารณะ

ขนาดชุดข้อมูลทั้งหมด (78.48GB)

■ wisights-large ■ pantip-large ■ ชุดข้อมูลเปิดทั้งหมด



เทรนโมเดลที่ใหญ่ที่สุดในประวัติศาสตร์ NLP ไทย

เราเลือกใช้สถาปัตยกรรม RoBERTa ที่สามารถเทรนได้ด้วยงาน masked language model (MLM) เท่านั้น ไม่เหมือน BERT ที่ต้องทำการทนายประโยคต่อท้าย (sentence entailment) ด้วย แต่การจะเทรนโมเดลที่ใหญ่ที่สุดเท่าที่เคยถูกเทรนบนภาษาไทยภาษาเดียวให้ได้นั้น มันไม่ได้ง่ายขนาดแคกดาวน์โหลดตาม tutorial ของ HuggingFace (แต่มันก็เป็น tutorial ที่ดีมาก เราแนะนำให้ไปลองอ่าน)

ปัญหาสำคัญคือ **ความแตกต่างในเชิงทรัพยากรการคำนวณ** ในขณะที่เราเทรนโมเดลด้วย Nvidia DGX-1 (ราคาตลาด 129,000 ดอลลาร์สหรัฐหรือประมาณเครื่องละ 4 ล้านบาท) ซึ่งประกอบด้วย GPU รุ่น Nvidia Tesla V100 ขนาด 32GB จำนวน 8 หน่วย ใช้เวลาประมาณ 125 วัน 19 ชั่วโมงต่อหนึ่งรอบการเทรนให้ครบ

โมเดลและคลังข้อมูลสาธารณะ



PyThaiNLP: Thai Natural Language Processing in Python

[pypy v3.0.5](#)
[python 3.7](#)
[license Apache 2.0](#)
[downloads/month 57k](#)
[Unit test and code coverage passing](#)

[coverage 97%](#)
[code quality A](#)
[license scan failing](#)
[Launch Quick Start Guide](#)
[on Google Colab](#)

[DOI 10.5281/zenodo.6075269](#)
[matrix join chat](#)

PyThaiNLP is a Python package for text processing and linguistic analysis, similar to [NLTK](#) with focus on Thai language.

PyThaiNLP เป็นไลบรารีภาษาไพทอนสำหรับประมวลผลภาษาธรรมชาติ คล้ายกับ NLTK โดยเน้นภาษาไทย ดูรายละเอียดภาษาไทยได้ที่ [README_TH.MD](#)

News

Now, You can contact or ask any questions you encounter with the PyThaiNLP team.

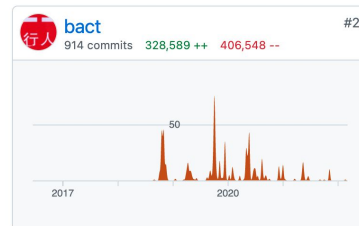
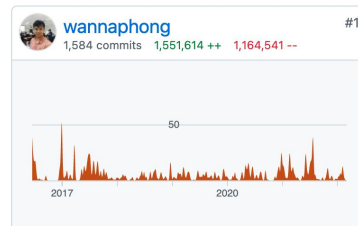
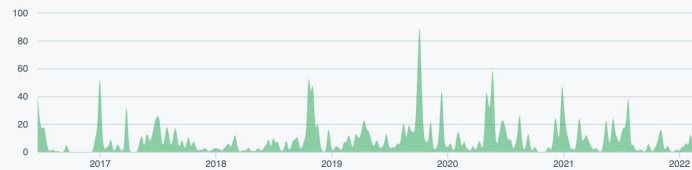
matrix join chat

Version	Description	Status
3.0	Stable	Change Log
dev	Release Candidate for 3.1	Change Log

Jun 19, 2016 – Mar 6, 2022

Contributions: Commits ▼

Contributions to dev, excluding merge commits and bot accounts



โมเดลและคลังข้อมูลสาธารณะ



ชุดข้อมูลและโมเดล ติดต่อเรา

ชุดข้อมูลและโมเดลที่น่าสนใจ

Thai Speech Emotion Dataset

ชุดข้อมูลจำแนกอารมณ์จากเสียงพูดภาษาไทย

26 Mar 2021

dataset v1.0

WangchanBERTa: Pre-trained Thai Language Model

โมเดลภาษาสำหรับงานประมวลผล และการเข้าใจภาษาไทย

3 Mar 2021

model v1.0

English-Thai Machine Translation Models

โมเดลแปลภาษา อังกฤษ-ไทย จากชุดข้อมูลกว่า 1 ล้าน คู่ประโยค

23 Jun 2020

model demo v1.0


English-Thai Machine Translation Dataset

ชุดข้อมูลคู่ประโยคภาษาอังกฤษ-ไทย กว่า 1 ล้านคู่ประโยค

23 Jun 2020

dataset v1.0

โมเดลและคลังข้อมูลสาธารณะ

 NLP For Thai

NLP For Thai

Other

TASKS

Thai NLP Tasks

Dependency Parser

Grapheme to Phoneme

Image Captioning

Language model

Machine Translation

Named Entity Recognition

Natural Language Inference

Optical Character Recognition

Part-of-speech tagging

Plagiarism

Question Answering

Sentence Segmentation

Speech Emotion Recognition

Spoken Language Understanding

Soundex

Speech Recognition

Speech Synthesis

Spell Correct

Syllable Segmentation

Text Classification

Text Generation

Text Summarization

Treebank

Word Segmentation

Docs » NLP For Thai

NLP For Thai



It's Thai NLP homepage. All is Open Source.

Website: [NLPForThai.com](https://nlpforthai.com) maintained by PyThaiNLP

Menu

- Tasks
- Other

Contributors

Thanks all the [contributors](#). (Image made with [contributors-img](#))

How to Contribute

You can fork and send your pull request at <https://github.com/PyThaiNLP/nlpforthai.com>

We build Thai NLP.

PyThaiNLP

Next

Built with [MkDocs](#) using a [theme](#) provided by [Read the Docs](#).

อะไรที่ภาษาไทยยังทำไม่ได้?

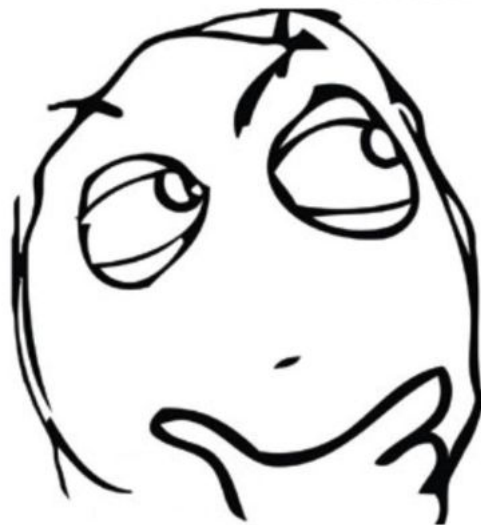


อะไรที่ภาษาไทยยังทำไม่ได้?

ตัดคำ

Method	WS160		TNHC		LST20		VISTEC	
	Char	Word	Char	Word	Char	Word	Char	Word
DC	93.47	84.03	89.48	75.40	94.60	87.15	92.77	81.78
AC	93.50	84.04	88.82	73.71	95.24	87.21	91.47	79.31
TL-DC	96.30	90.60	95.43	88.60	98.63	96.30	96.78	90.99
TL-AC	94.10	85.00	90.57	77.54	98.04	94.77	95.47	89.27
SE-DC	95.20	86.90	95.20	84.10	94.96	87.72	94.76	86.33
SE-AC	94.50	85.60	93.70	83.90	96.30	89.87	93.86	84.43
DSE-DC	96.67	91.51	95.71	89.14	99.01	97.33	97.36	92.91
DSE-AC	94.57	86.24	95.51	88.52	98.46	95.79	97.31	92.78

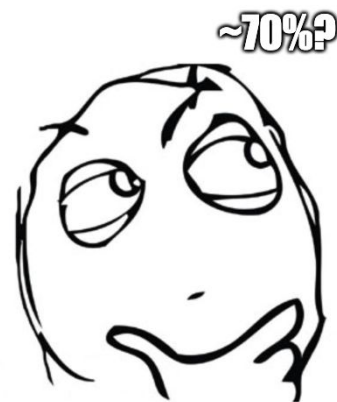
90%?



อะไรที่ภาษาไทยยังทำไม่ได้?

การจำแนกข้อความ

	โมเดล-ชุดข้อมูล	Wisesight Sentiment	Wongnai Reviews	Generated Reviews EN- TH	Prachathai67k
1	NBSVM	72.03	58.38	59.68	66.77
2	ULMFit	70.95	61.79	64.33	66.21
3	XLMR	73.57	62.57	64.91	68.18
4	mBERT	70.05	47.99	62.14	66.47
5	WanchanBERTa	76.19	63.05	64.66	69.78

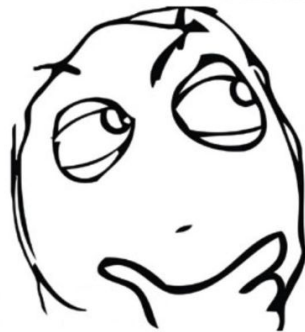


อะไรที่ภาษาไทยยังทำไม่ได้?

เครื่องแปลภาษา

	Google	AI for Thai	Our Baseline (SCB_1M)	Our Baseline (MT_OPUS)	Our Baseline (SCB_1M+MT_OPUS)
<i>Thai → English</i> IWSLT 2015					
SacreBLEU (case-sensitive)	14.19 (46.7/19.9/10.0/5.1)	※	17.2 (50.7/23.1/12.1/6.6)	28.1 (60.8/35.6/23.1/15.)	28.3 (60.8/35.6/22.9/15.1)
SacreBLEU (case-insensitive)	17.64 (53.8/24.5/12.7/6.8)	※	17.93 (52.4/24.0/12.7/7.0)	28.7 (62.0/36.3/23.7/16.)	29.0 (62.0/36.4/23.5/15.6)
<i>English → Thai</i> IWSLT 2015					
BLEU4	15.36 (51.0/23.8/12.0/6.2)	6.14 (36.1/11.7/4.3/1.7)	12.95 (45.5/19.5/9.0/4.3)	17.24 (52.0/26.3/14.4/8.1)	17.77 (52.2/26.7/14.8/8.5)

ดีกว่า GOOGLE?



อะไรที่ภาษาไทยยังทำไม่ได้?

ทำตัดคำไม่ได้



NLP ในยุคต้น 2000

มีโมเดลเทพพอที่
จะไม่ต้องแคร์ผล
การตัดคำ



NLP ในยุค 2020

เริ่มทำ NLP เริ่มยังไงดี?

Thai

Natural Language Processing

กลุ่มคนทำ NLP ภาษาไทย

Thai Natural Language Processing

👤 กลุ่มสาธารณะ · สมาชิก 1.2 หมื่น คน

👤 ได้เข้าร่วม ▼

+ เชิญ

▼

☰ README.md



PyThaiNLP: Thai Natural Language Processing in Python

[pypi v3.0.5](#)
[python 3.7](#)
[License Apache 2.0](#)
[downloads/month 57k](#)
[Unit test and code coverage passing](#)

[coverage 97%](#)
[code quality A](#)
[license scan failing](#)
[Launch Quick Start Guide on Google Colab](#)

[DOI 10.5281/zenodo.6075269](#)
[matrix join chat](#)

PyThaiNLP is a Python package for text processing and linguistic analysis, similar to [NLTK](#) with focus on Thai language.


PyThaiNLP เป็นไลบรารีภาษาไพทอนสำหรับประมวลผลภาษาธรรมชาติ คล้ายกับ NLTK โดยเน้นภาษาไทย 🇹🇹
รายละเอียดภาษาไทยได้ที่ [README_TH.MD](#)

News

Now, You can contact or ask any questions you encounter with the PyThaiNLP team.
[matrix join chat](#)

Version	Description	Status
3.0	Stable	Change Log
dev	Release Candidate for 3.1	Change Log


คอร์สแนะนำ



Natural Language Processing

วิดีโอ 15 รายการ • การดู 25,302 ครั้ง • อัปเดตล่าสุดเมื่อ 9 เม.ย. 2021


Course github:
https://github.com/ekapolc/nlp_course



Ekapol Chuangsuwanich


ติดตามแล้ว

- 1



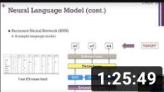
2110594 NLP L1 Introduction

Ekapol Chuangsuwanich
- 2




2110594 NLP L2 Tokenization: Intro to Neural Networks

Ekapol Chuangsuwanich
- 3




2110594 NLP L3 Language Modeling

Ekapol Chuangsuwanich
- 4



2110594 NLP L4 Word representation

Ekapol Chuangsuwanich
- 5



2110594 NLP L5 PoS Tagging

Ekapol Chuangsuwanich

คอร์สแนะนำ

Text Classification

โจทย์ส่วนใหญ่ของ NLP นั้นสามารถแก้ได้ด้วยการสร้างเครื่องจำแนกประเภทข้อความ (Text Classifier) ซึ่งมี machine learning algorithm เป็นแกนหลัก (สามารถศึกษาพื้นฐานของ machine learning จาก module [sentiment analysis](#)) ปัจจุบันนี้คนหันมาใช้ Neural Network กันมากขึ้น โดยใช้พื้นฐานของความหมายของคำที่กำหนดโดยบริบทและการใช้ภาษา (Distributed semantic model หรือ distributional semantic model) ในคลังข้อมูลเป็นหลัก



[Text Classification - NLP] 1 Text Classification คือ...

คุณงานชน|สื่อชน|แกะใส่|าระบะ

- Part of speech tagging : แปลชนิดของคำ
- Input: คำแต่ละคำ
- Output: {Noun, Adj, Verb, Adv, Preposition, ..}

TEXT CLASSIFICATION

Video List:

1. [Text Classification คืออะไร](#)
2. [Logistic Regression แบบไม่ต้องเขียนสมการ](#)
3. [Crossentropy Loss ของ Logistic Regression](#)
4. [วิธีการฝึกโมเดลด้วย scikit-learn](#)
5. [Sparse features and sparse matrix](#)
6. [Data structure: numpy array and dense matrix](#)
7. [Data structure: Sparse Matrix](#)

เนื้อหาสำหรับอ่านและอ้างอิง

- [Slides จากวิดีโอเรื่อง Text classification](#)
- [Chapter 4 จาก Jurafsky](#)
- [Chapter 5 จาก Jurafsky](#)



THANK YOU

