

การเทรนโมเดลบน Hugging Face พาร์ท 2

Peerat Limkonchotiwat

PhD student at VISTEC, Thailand

AGENDA

What We'll Learn Today

Machine Translation

Question answering

Representation Learning



สารบัญ

เราจะเรียนอะไรกันในวันนี้

เครื่องแปลภาษา (Machine Translation)

ระบบถาม-ตอบ (Question-Answering)

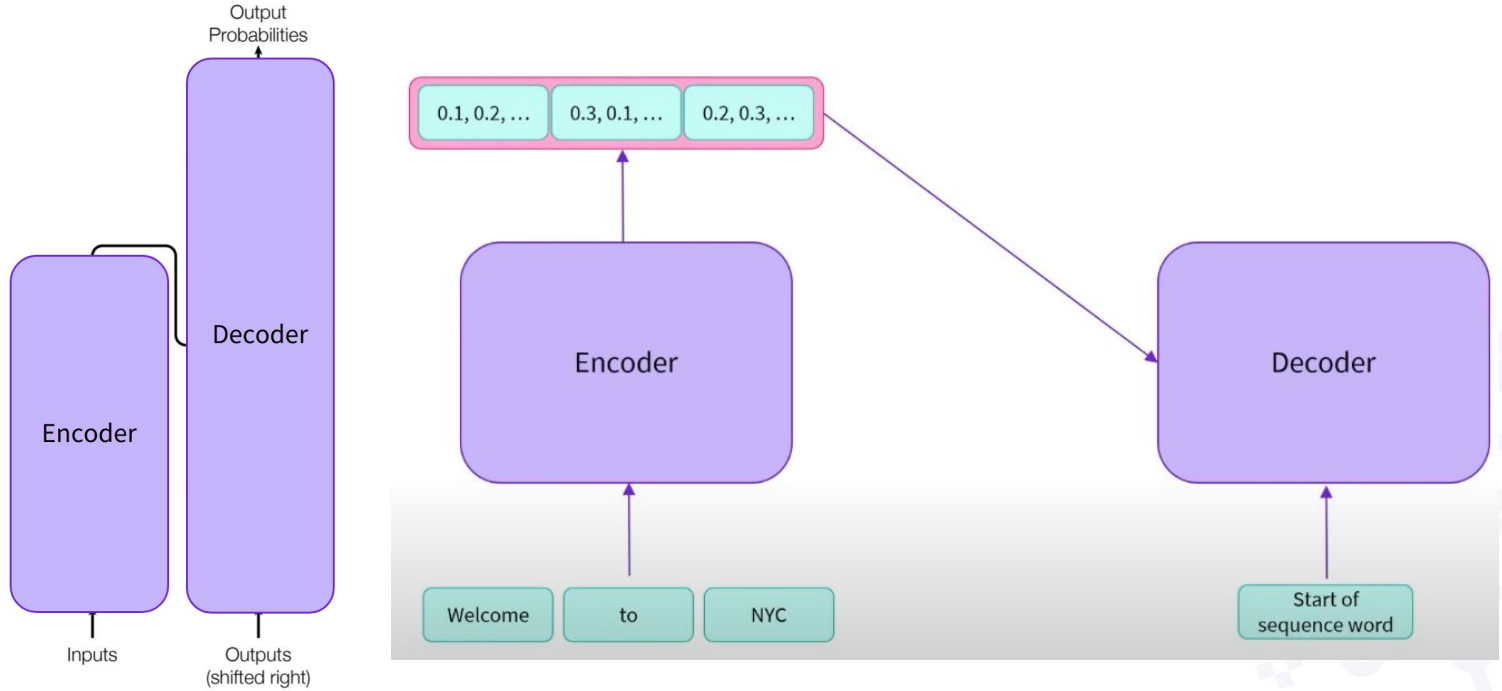
โมเดลแปลง text \Rightarrow Vector (Representation Learning)



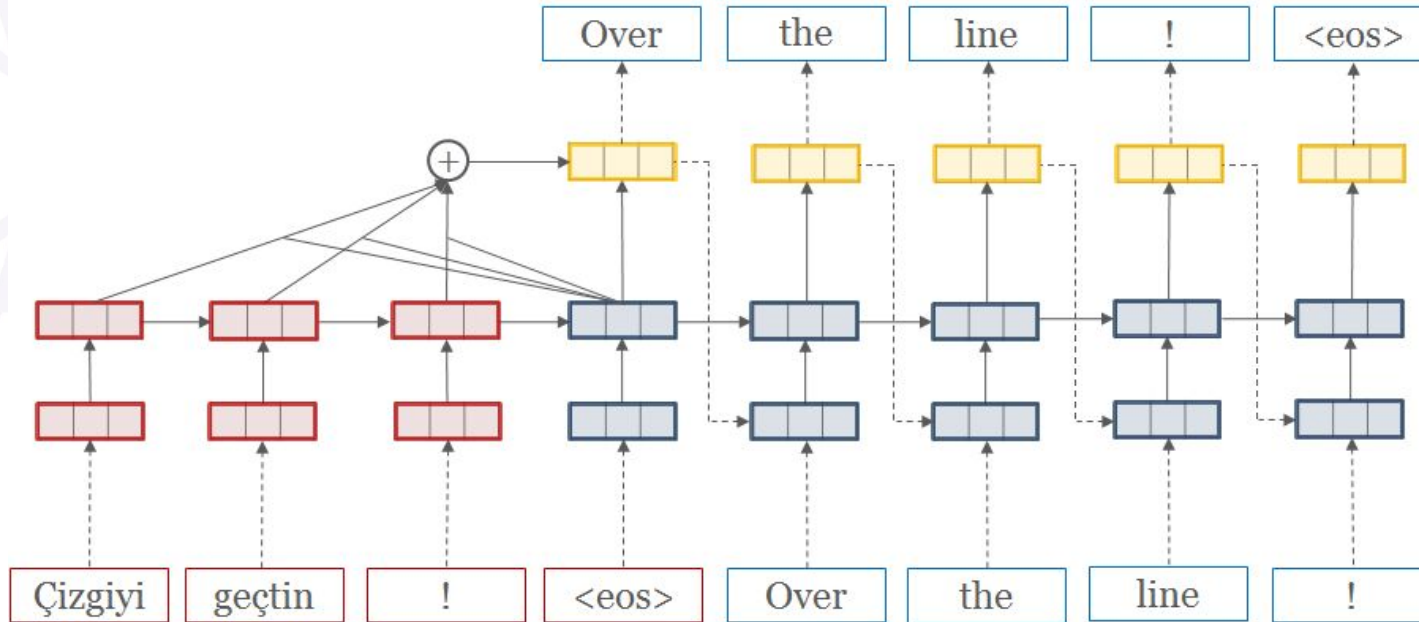


Topic 1: เครื่องแปลภาษา Machine Translation

Encoder-Decoder



Machine Translation



โมเดลของไทย

- Translator from Lan1 \Rightarrow Lan2
- Require parallel corpus, e.g.,
 - scb-mt-en-th-2020
 - English - Thai dataset
 - 1 M sentences

Sub-dataset	Method	Number of segment pairs
Taskmaster-1	Professional Translators	222,733
Product Reviews Translation		133,330
Product Reviews Annotation	Annotation by translators	280,208
NUS SMS Messages		43,750
Microsoft Research Paraphrase Identification	Crowd-sourced translators	10,371
Mozilla Common Voice		33,797
Product Reviews Translation		24,587
Government Documents	PDF documents	25,398
Top-500 Thai Websites		120,280
ParaCrawl	Web crawling	60,039
Wikipedia		33,756
Asia Pacific Defense Forum		13,503
		1,001,752

	Google	AI for Thai	Our Baseline (SCB_1M)	Our Baseline (MT_OPUS)	Our Baseline (SCB_1M+MT_OPUS)
<i>Thai \rightarrow English IWSLT 2015</i>					
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 \rightarrow Thai IWSLT 2015</i>					
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)

Machine Translation

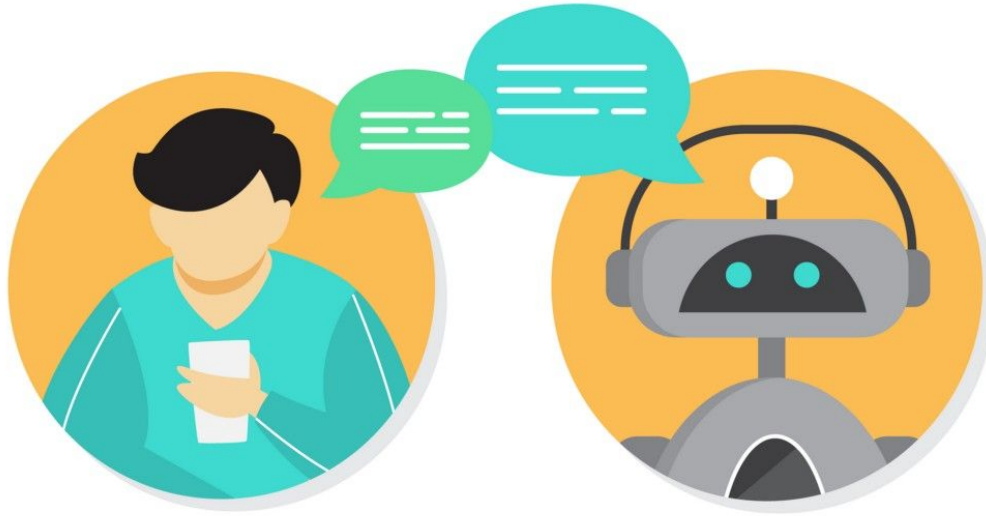
Let's code!





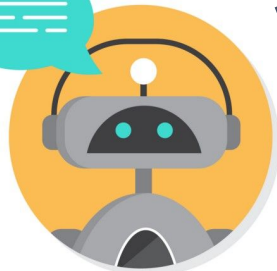
Topic 2: ระบบถามตอบ Question Answering

Question Answering

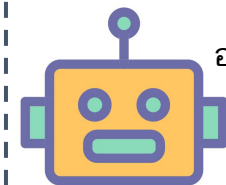


Question Answering

Question: ใครคือ
ประธานาธิบดีของสหรัฐใน
ปี 2022



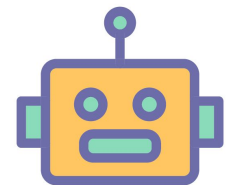
Question: โจ ไบเดน



อ่านคลังข้อมูล



เปิดหนังสือ หรือ Open book



ฉันมีข้อมูลอยู่ในหัว
แล้ว!!

ปิดหนังสือ หรือ Close book/Open domain

Question Answering

Let's code!





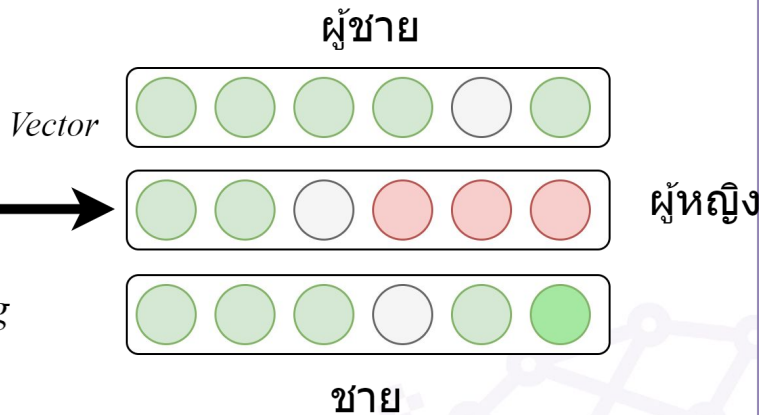
Topic 3: Representation Learning

อะไรคือ Rerepresentation บน NLP?

Text Dictionary Lookup

ผู้ชาย	1
ผู้หญิง	2
ชาย	3

Algorithm Encoder
Machine learning
Deep learning
or bag-of-word

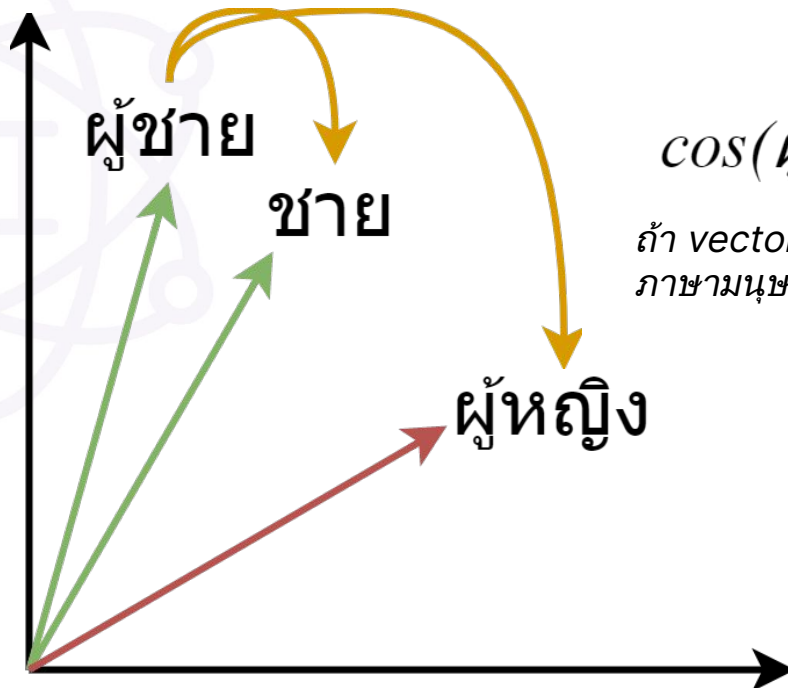


ฉันไม่เข้าใจตัวอักษร



แล้วเลข 1,2,3 มีความ
ยังไง?

เพราะฉนั้น?



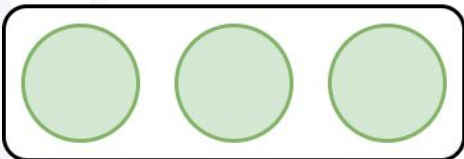
Cosine similarity ($\cos(\bullet)$)

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

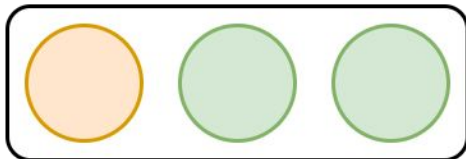
ถ้า vector ของสามารถแปลงความหมายได้ดี = คอมพิวเตอร์เข้าใจภาษามนุษย์ได้ดี

สร้าง Sentence Embedding จาก Word

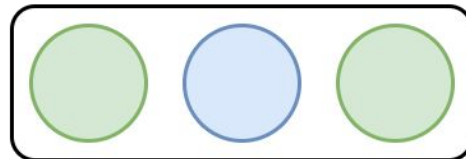
กิน



ข้าว

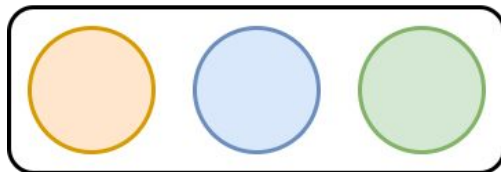


กัน

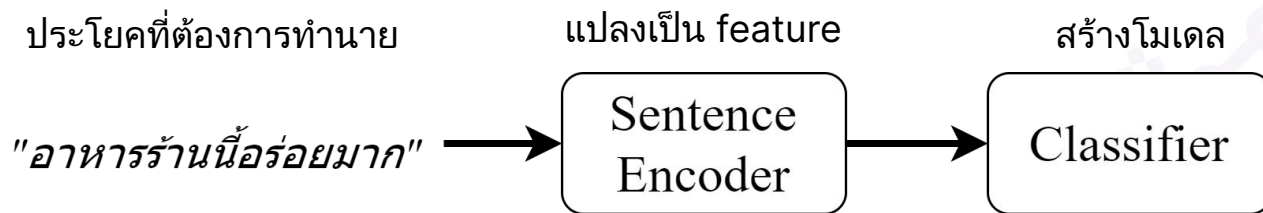
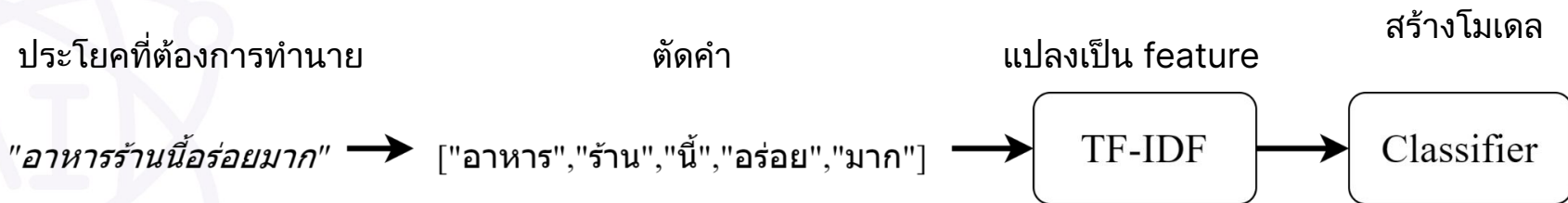


Mean

กินข้าวกัน



เอามาใช้กับ Text Classification?



ไม่ต้องเขียน feature อะไรให้ยุ่งยาก, ใช้แค่ Encoder!!

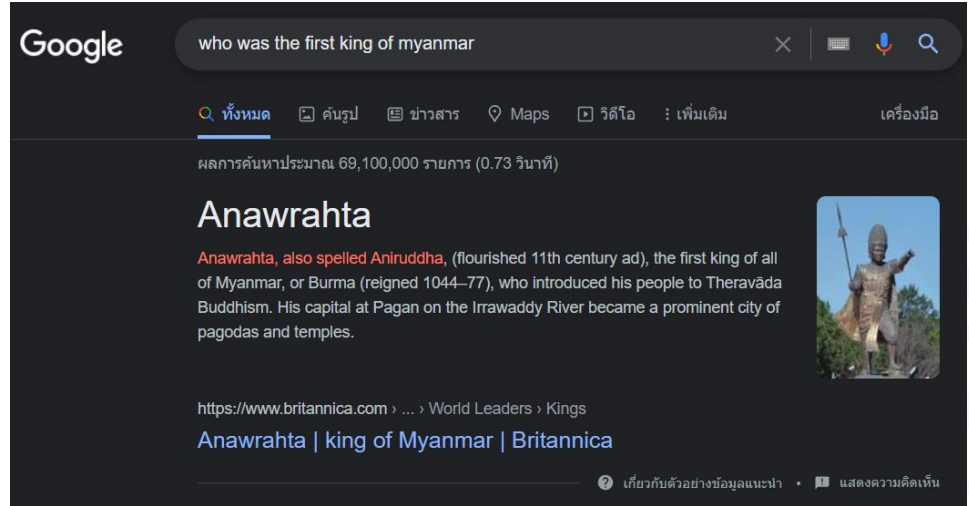
การเอาไปใช้งานด้านอื่นๆ

Query: Who was the first king of Myanmar



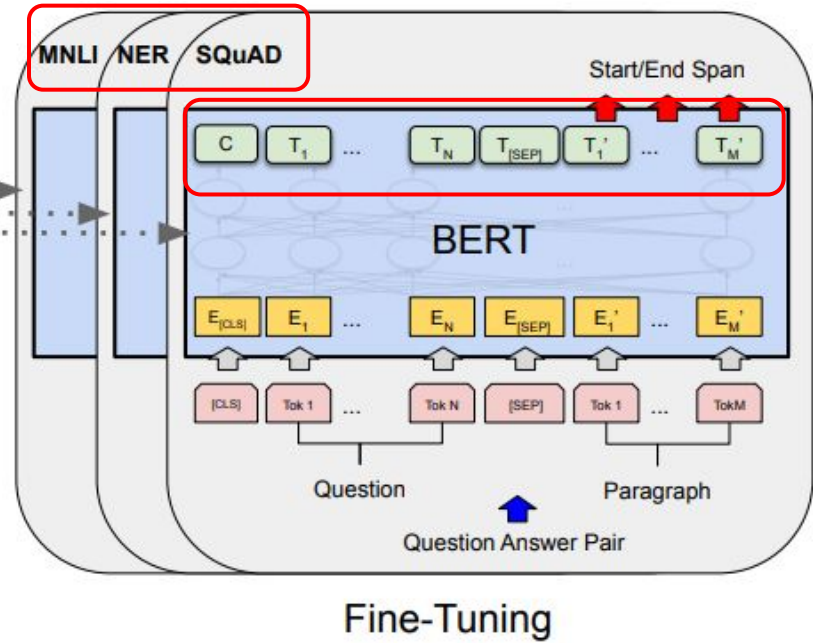
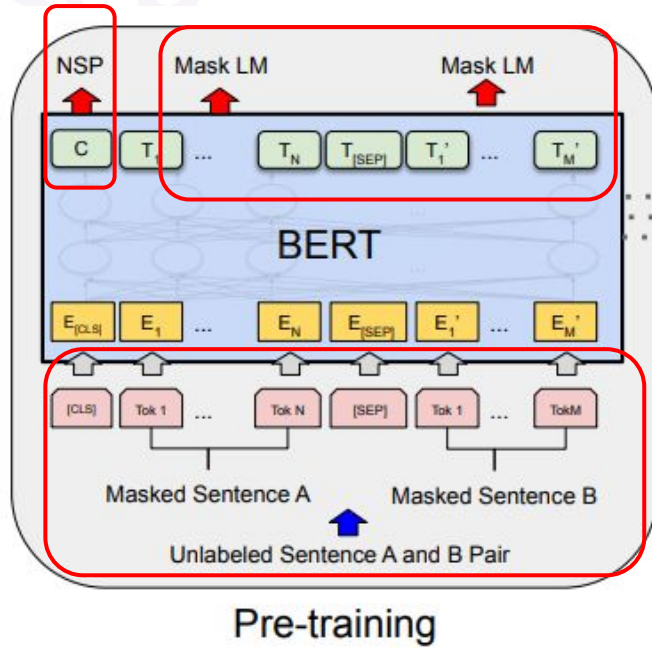
*Document
search*

WIKIPEDIA
The Free Encyclopedia

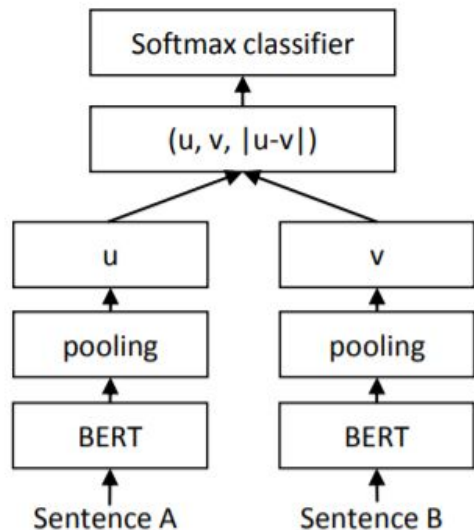
Google search results for "who was the first king of myanmar". The search bar shows the query. Below the search bar, there are navigation options: ทั้งหมด, ค้นรูป, ข่าวสาร, Maps, วิดีโอ, and เพิ่มเดิม. The search results show approximately 69,100,000 results in 0.73 seconds. The main result is for "Anawrahta", described as the first king of Myanmar, who introduced Theravāda Buddhism. A small image of a statue of Anawrahta is shown. The result is from Britannica, with the URL <https://www.britannica.com>.

Sentence representation จาก BERT

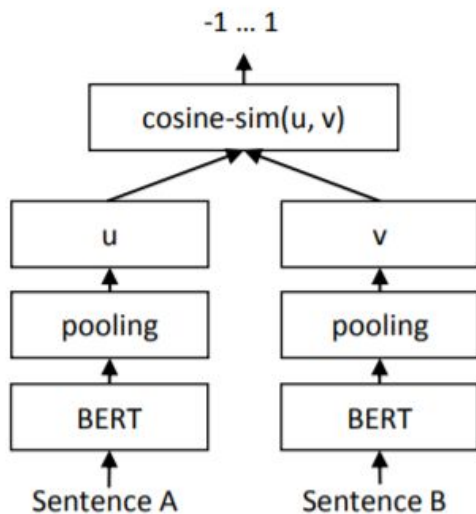


ยกตัวอย่าง: Sentence-transformer

Language Understanding



Semantic Understanding

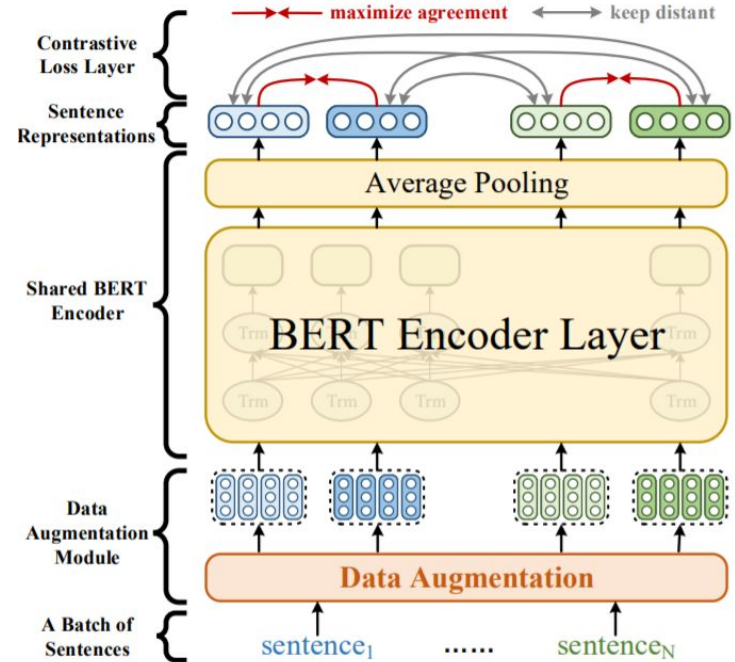
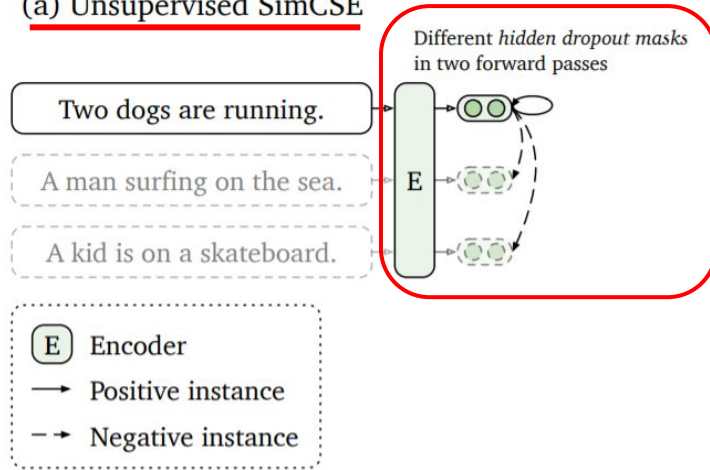


Model	Spearman
<i>Not trained for STS</i>	
Avg. GloVe embeddings	58.02
Avg. BERT embeddings	46.35
InferSent - GloVe	68.03
Universal Sentence Encoder	74.92
SBERT-NLI-base	77.03
SBERT-NLI-large	79.23
<i>Trained on STS benchmark dataset</i>	
BERT-STSB-base	84.30 ± 0.76
SBERT-STSB-base	84.67 ± 0.19
SRoBERTa-STSB-base	84.92 ± 0.34
BERT-STSB-large	85.64 ± 0.81
SBERT-STSB-large	84.45 ± 0.43
SRoBERTa-STSB-large	85.02 ± 0.76
<i>Trained on NLI data + STS benchmark data</i>	
BERT-NLI-STSB-base	88.33 ± 0.19
SBERT-NLI-STSB-base	85.35 ± 0.17
SRoBERTa-NLI-STSB-base	84.79 ± 0.38
BERT-NLI-STSB-large	88.77 ± 0.46
SBERT-NLI-STSB-large	86.10 ± 0.13
SRoBERTa-NLI-STSB-large	86.15 ± 0.35

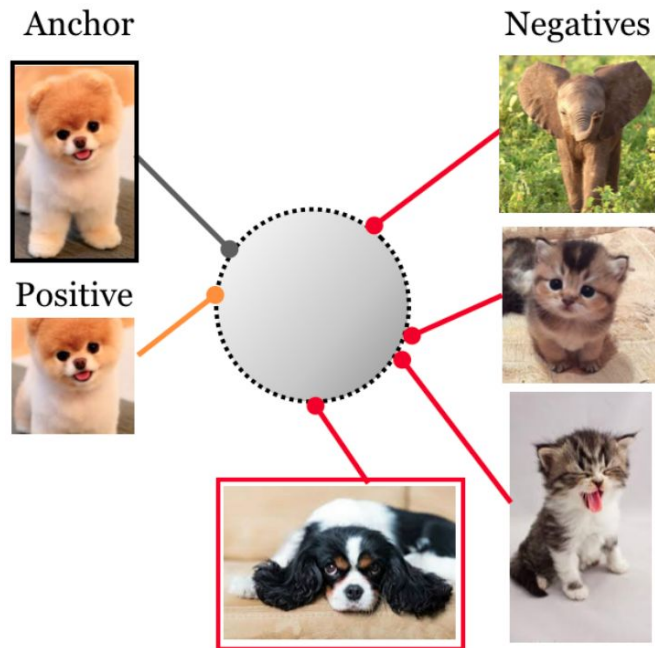
วิธียอดนิยม: Contrastive Learning

$$l_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_i^{z'_i})/\tau}}{\sum_{j=1}^N e^{\text{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_j^{z'_j})/\tau}},$$

(a) Unsupervised SimCSE



อะไรคือ Contrastive Learning?



Representation Learning

Let's code!





THANK YOU

