AI Workshop: Self-supervised learning models in NLP

Introduction to BERT and its family

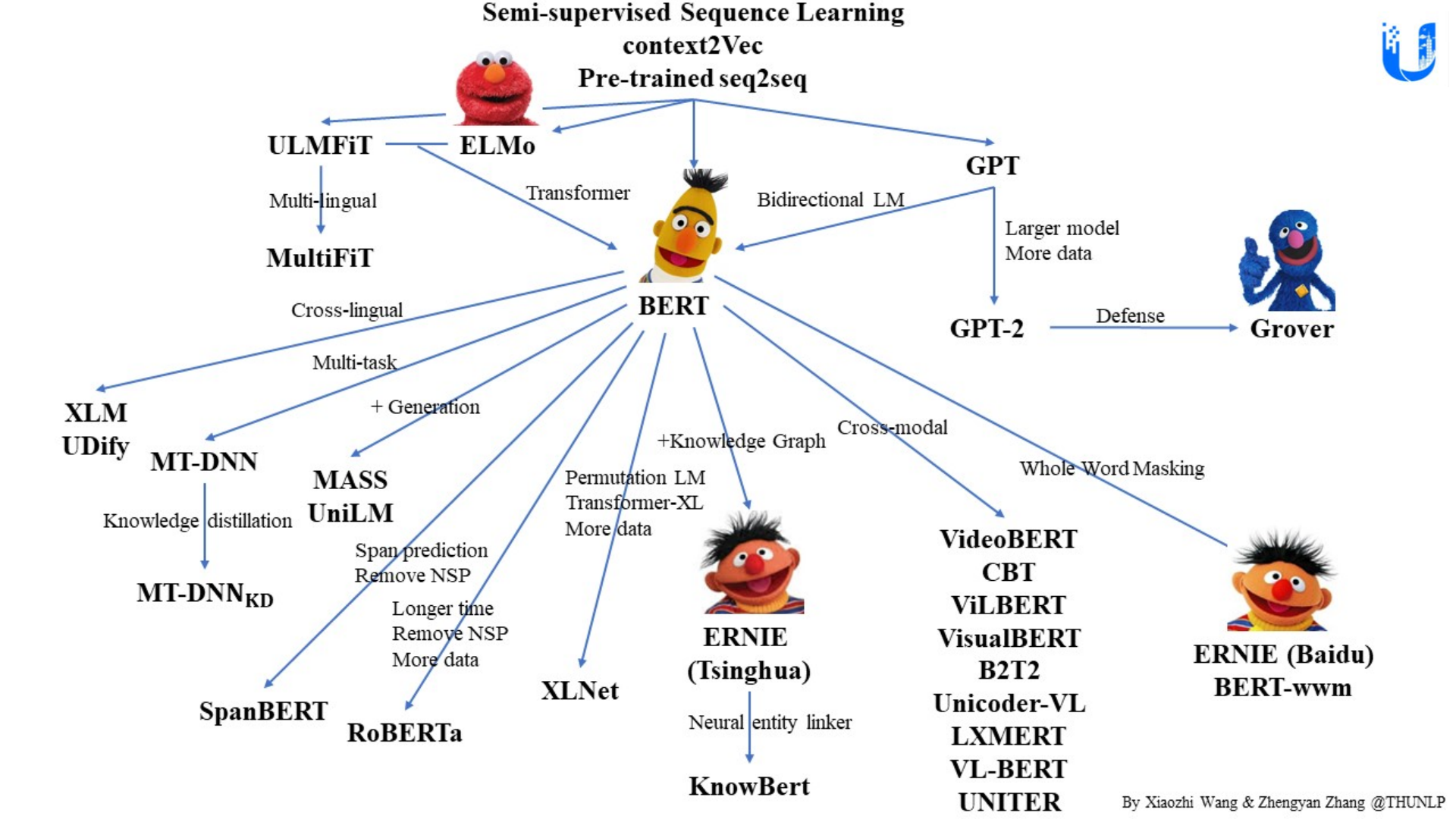
Present@2022/03/03 Leo

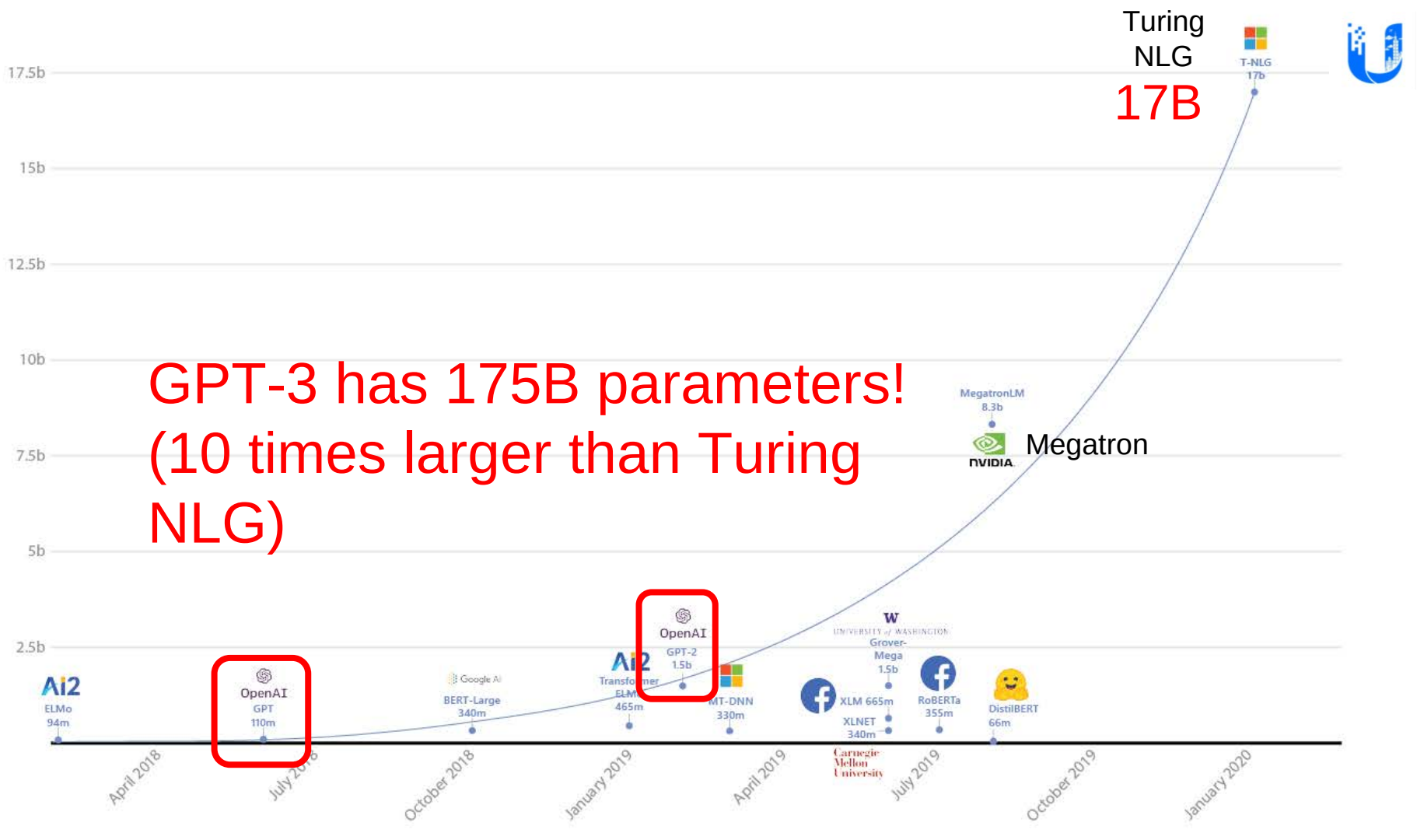
Slido:
#BERTology_Leo
Feedback:
https://forms.gle/YGcdxwjcma6LP3iK6

Slide Download in slido

Outline

1. Contextualized embedding & Self-supervised learning
   A. BERT Introduction
   B. How to use BERT
   C. Why does BERT work
2. BERT & its family/relatives
   A. GPT
   B. RoBERTa
   C. SpanBERT
   D. ALBERT
3. Multi-lingual support & Cross-lingual language models
   A. mBERT
   B. XLM
4. Q & A


Semi-supervised Sequence Learning
context2Vec
Pre-trained seq2seq
ULMFiT
ELMo
GPT
Multi-lingual
MultiFiT
Transformer
Bidirectional LM
Larger model
More data
BERT
GPT-2
Defense
Grover
Cross-lingual
XLM
UDify
Multi-task
MT-DNN
Knowledge distillation
MT-DNN$_{KD}$
+ Generation
MASS
UniLM
Span prediction
Remove NSP
SpanBERT
Longer time
Remove NSP
More data
RoBERTa
Permutation LM
Transformer-XL
More data
XLNet
+Knowledge Graph
ERNIE
(Tsinghua)
Neural entity linker
KnowBert
Cross-modal
VideoBERT
CBT
ViLBERT
VisualBERT
B2T2
Unicoder-VL
LXMERT
VL-BERT
UNITER
Whole Word Masking
ERNIE (Baidu)
BERT-wwm
By Xiaozhi Wang & Zhengyan Zhang @THUNLP


Turing
NLG
17B
T-NLG
17b
17.5b
15b
12.5b
10b
7.5b
5b
2.5b
GPT-3 has 175B parameters!
(10 times larger than Turing NLG)
MegatronLM
8.3b
NVIDIA
Megatron
Ai2
ELMo
94m
OpenAI
GPT
110m
Google AI
BERT-Large
340m
Ai2
Transformer
ELMo
465m
OpenAI
GPT-2
1.5b
MT-DNN
330m
XLM 665m
XLNET
340m
Carnegie
Mellon
University
UNIVERSITY of WASHINGTON
Grover-
Mega
1.5b
RoBERTa
355m
DistilBERT
66m
April 2018
July 2018
October 2018
January 2019
April 2019
July 2019
October 2019
January 2020

ELMo (94M)

BERT (340M)

GPT-3 (175B)

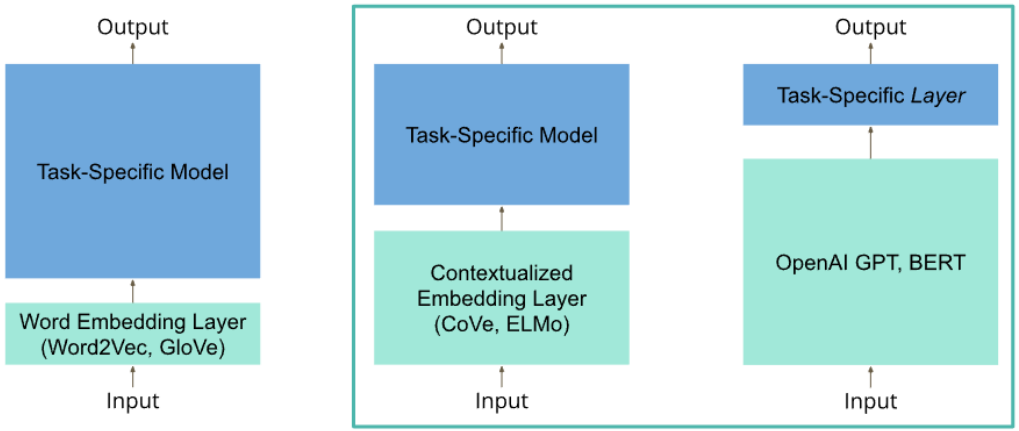
Switch Transformer (1.6T)
比玉山還要高

Sesame Street Language Models

Contextualized embedding & Self-supervised learning

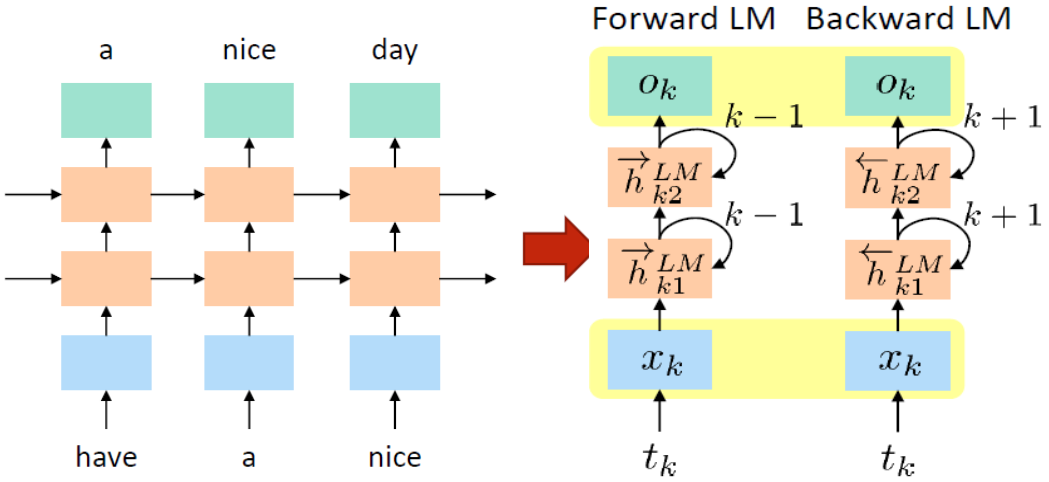
Strategies for applying pre-trained language representations

- Feature-based approach
  - ELMo (Peters et al., 2018)
  - Uses tasks-specific architectures that include the pre-trained representations as additional features
- Fine-tuning approach
  - Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018)
  - Minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning the pretrained parameters

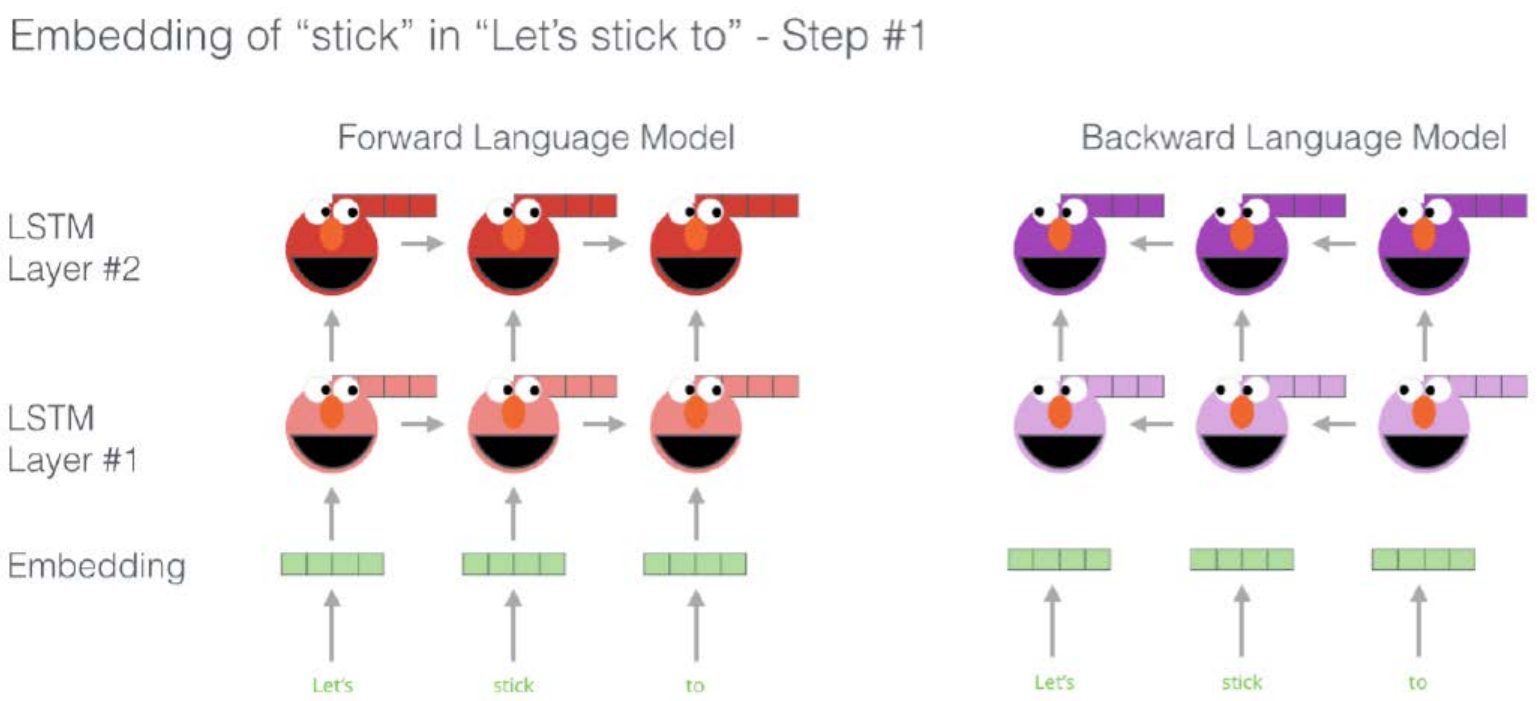


ELMo: Embeddings from Language Models

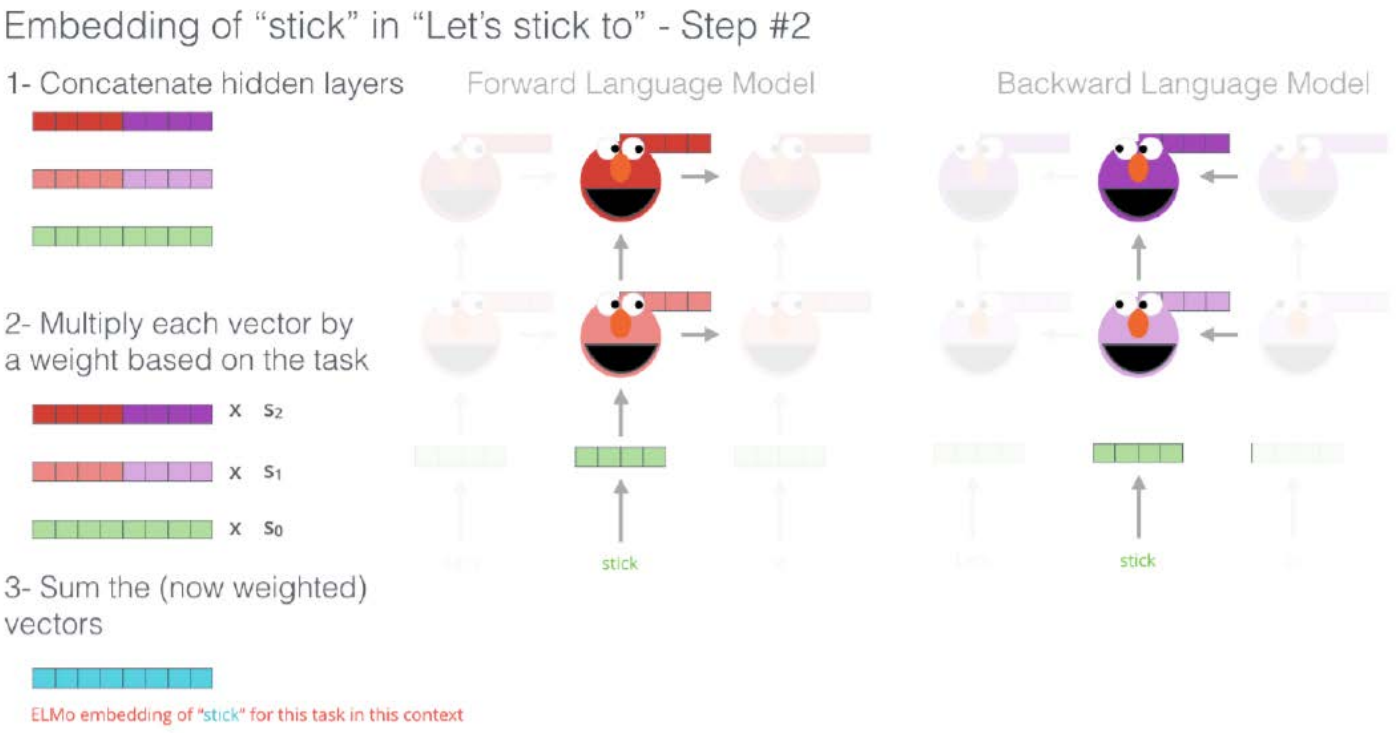
- A bidirectional language model
  - extract context sensitive features from a language model
- The contextualized representation is the concatenation of the output of the forward and backward LSTM
  - concatenation of independently trained left-to-right(LTR) and right-to-left(RTL) LSTM to generate features for downstream tasks



ELMo: Embeddings from Language Models

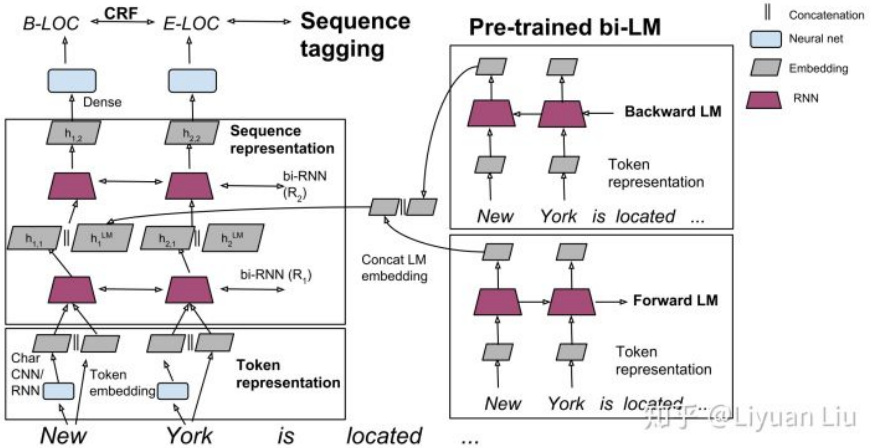


ELMo: Embeddings from Language Models



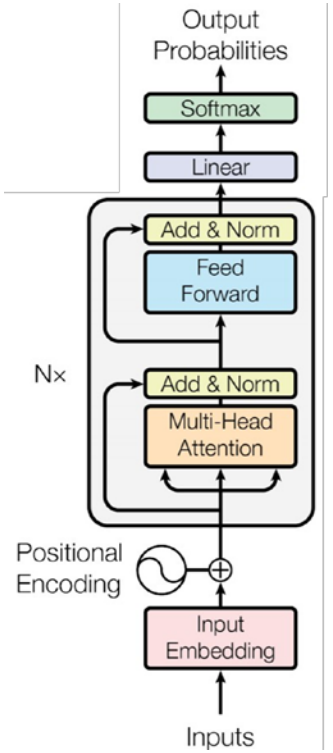
ELMo: Embeddings from Language Models

- Contextualized embeddings learned from LM provide informative cues
- ELMo – a general approach for learning high-quality deep context-dependent representations from biLMs
  - Pre-trained ELMo: https://allennlp.org/elmo
  - ELMo can process the character-level inputs



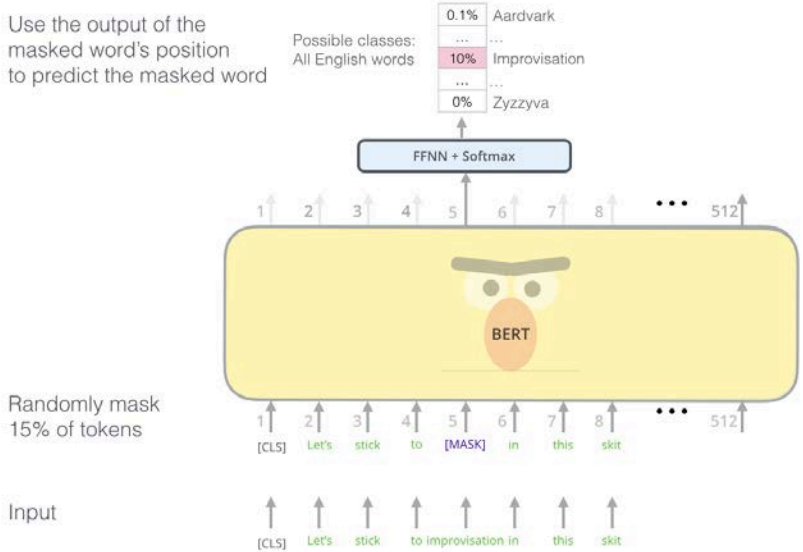
BERT: Bidirectional Encoder Representation from Transformers

- Idea: contextualized word representations
  - BERT is a multi-layer bidirectional Transformer encoder
  - Learning word vectors using long contexts using Transformer instead of LSTM
- It demonstrate the importance of bidirectional pre-training for language representations
  - Masked language models(MLM) to enable pre-trained deep bidirectional representations
- It is the first fine-tuning based representation model that achieves state-of-the-art performance on a large suite of sentence-level and token-level tasks
  - Outperforming many systems with task-specific architectures



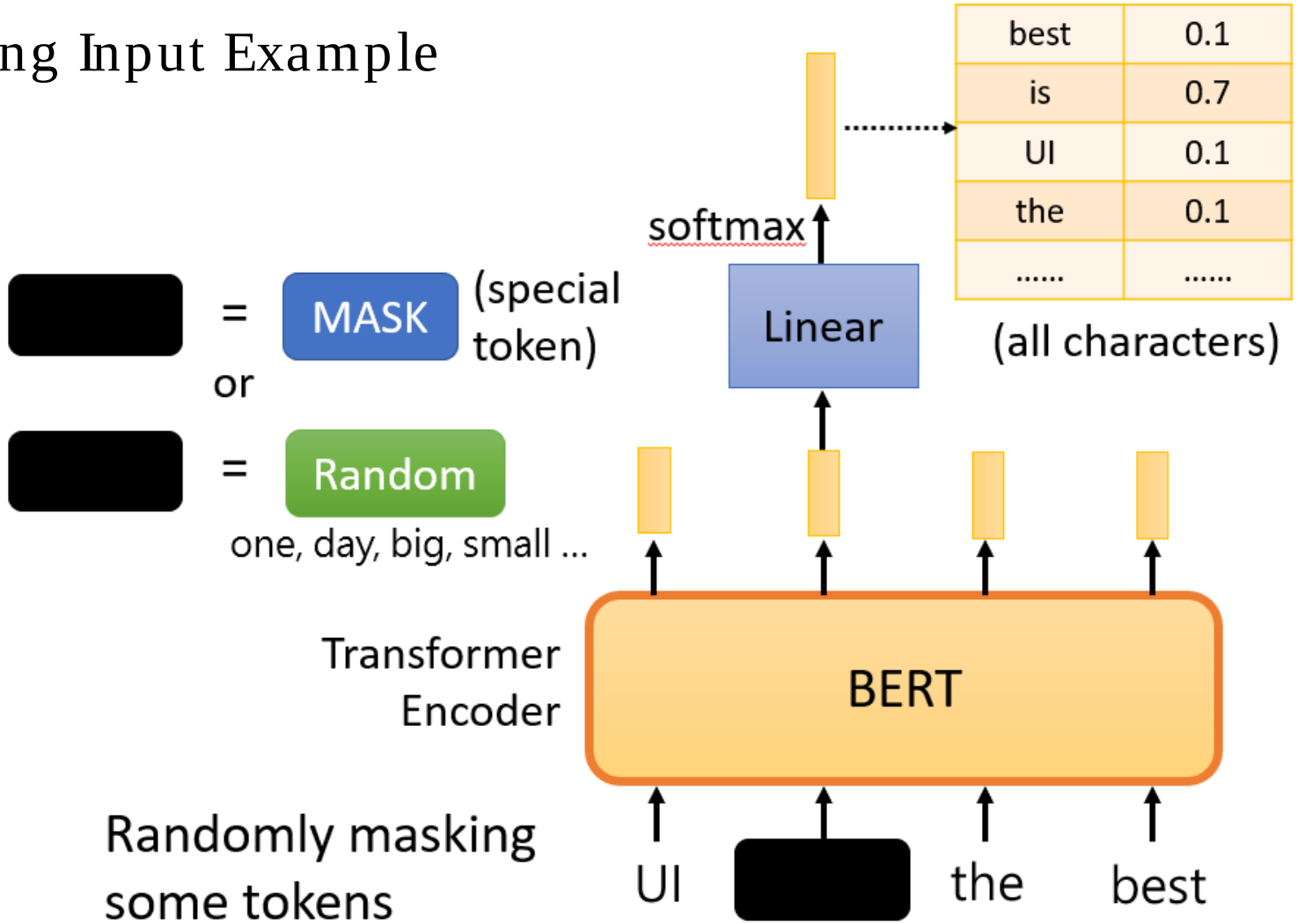
Task#1 – Masked Language Model (MLM)

- Idea: language understanding is bidirectional while LM only uses left or right context
  - Masking some percentage of the input tokens at random, and then predicting only those masked tokens
- 最後mask tokens的hidden vectors會輸出至vocabulary size的softmax去預測

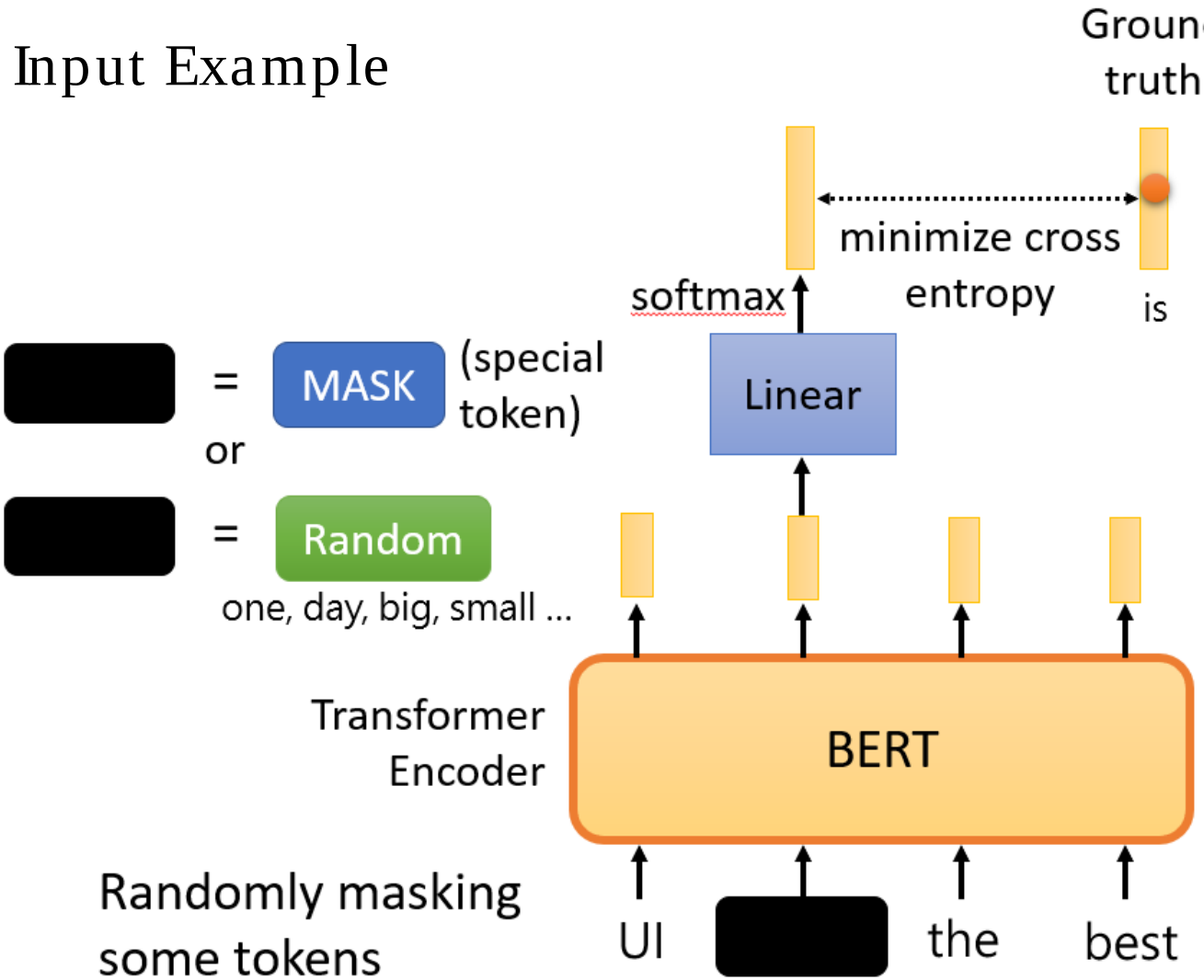


Randomly mask 15% of tokens
- Too little: expensive to train
- Too much: not enough context

Masking Input Example

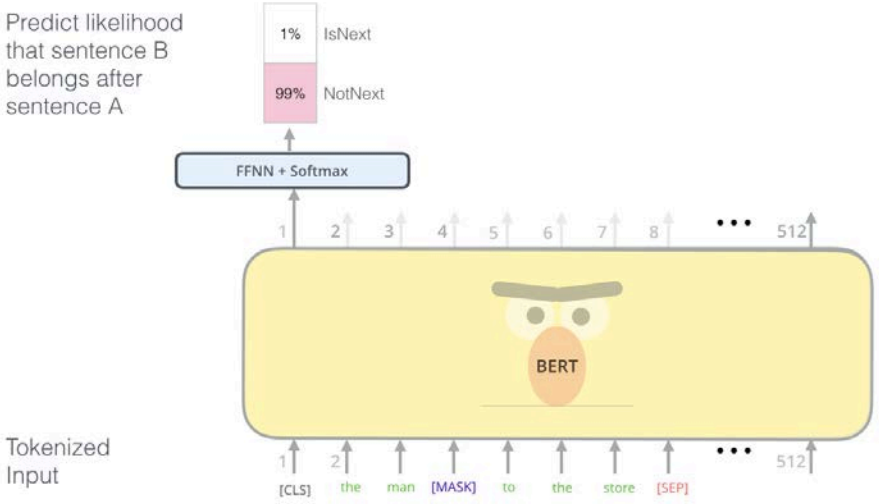


Masking Input Example



Task#2 – Next Sentence Prediction (NSP)

- Idea: To model relationships between sentences
  - jointly pre-trains text-pair representations
- 很多downstream task例如Question Answering (QA)、Natural Language Inference (NLI)是基於理解兩句話的關係的(inter-sentence relationship)，因此希望可以把相鄰上下兩句話的關係捕捉起來

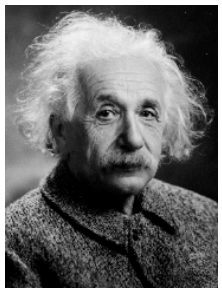


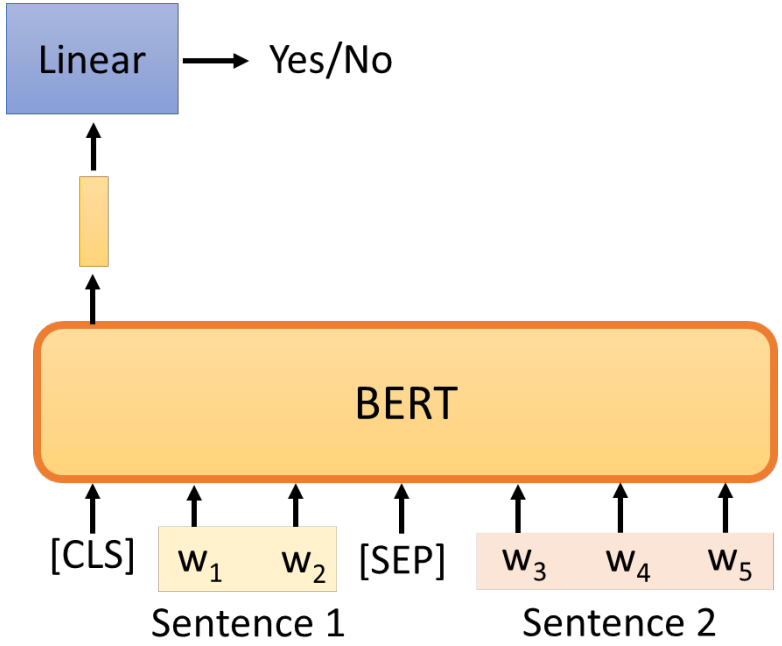
Input = [CLS] the man went to [MASK] store [SEP]
he bought a gallon [MASK] milk [SEP]
Label = IsNext

Input = [CLS] the man [MASK] to the store [SEP]
penguin [MASK] are flight ##less birds [SEP]
Label = NotNext

Task#2 – Next Sentence Prediction (NSP)

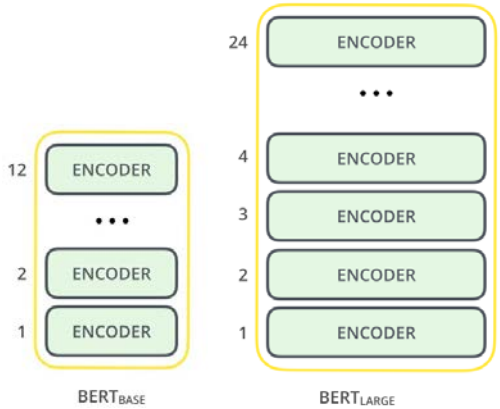
- This approach may not be helpful.
  - Robustly optimized BERT approach (RoBERTa)
- **SOP**: **S**entence **o**rder **p**rediction
  - Used in ALBERT

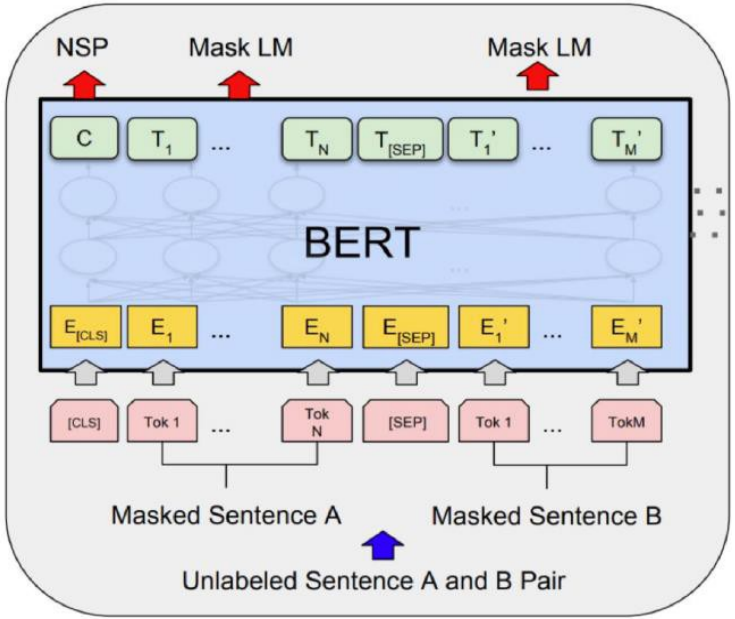




Pre-Training BERT

- Training data: Wikipedia + BookCorpus
- 2 BERT models
  - BERT-Base: 12-layer, 768-hidden, 12-head
  - BERT-Large: 24-layer, 1024-hidden, 16-head



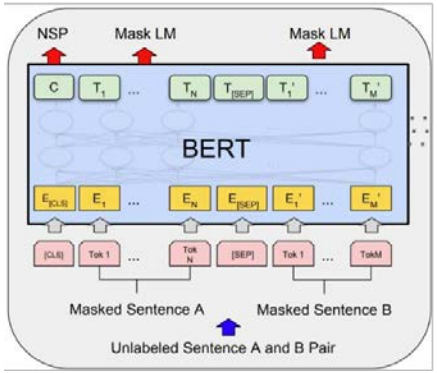


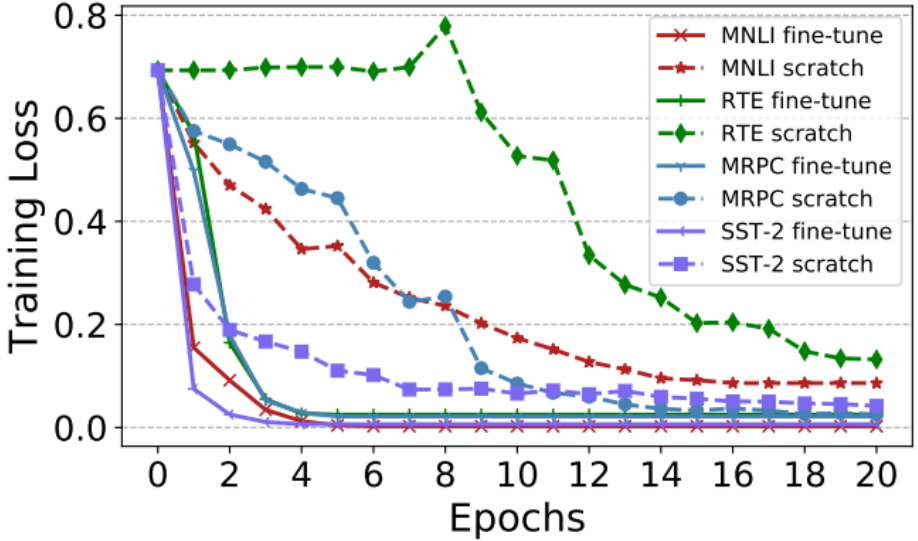
Applications of BERT

Downstream task finetuning

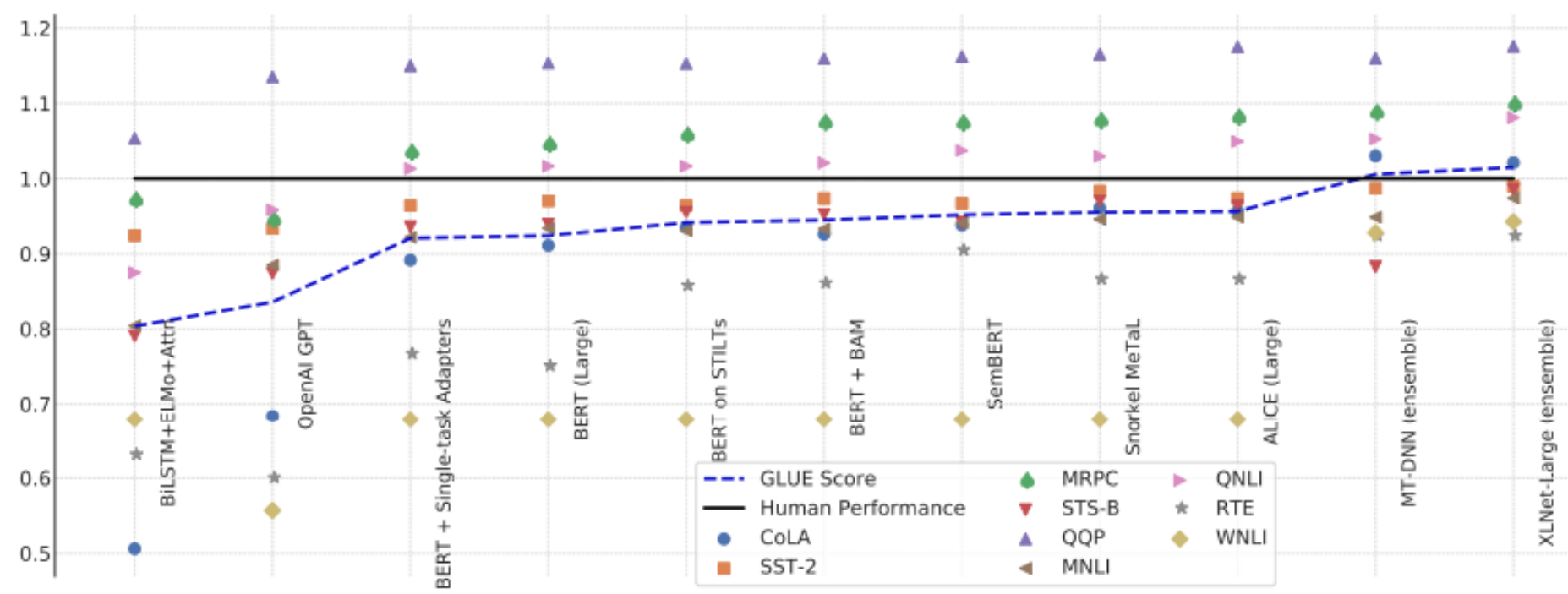
Fine-Tune BERT

- Idea: simply learn a classifier/tagger built on the top layer for each target task (BERT with one additional output layer)
  - minimal number of parameters need to be learned from scratch
  - sequence-level tasks / token-level tasks





GLUE: General Language Understanding Evaluation



https://gluebenchmark.com/
GLUE also has Chinese version (https://www.cluebenchmarks.com/)

MNLI

- Multi-Genre Natural Language Inference
- Given a pair of sentences, the goal is to predict whether the second sentence is an entailment(支持), contradiction(反對), or neutral(沒關係) with respect to the first one

Premise	Label	Hypothesis
Fiction		
The Old One always comforted Ca'daan, except today.	*neutral*	Ca'daan knew the Old One very well.
Letters		
Your gift is appreciated by each and every student who will benefit from your generosity.	*neutral*	Hundreds of students will benefit from your generosity.
Telephone Speech		
yes now you know if if everybody like in August when everybody's on vacation or something we can dress a little more casual or	*contradiction*	August is a black out month for vacations in the company.
9/11 Report		
At the other end of Pennsylvania Avenue, people began to line up for a White House tour.	*entailment*	People formed a line at the end of Pennsylvania Avenue.

QNLI

- Question Natural Language Inference (Stanford Question Answering Dataset)
- binary classification task
  - positive examples: (question, sentence) pairs which do contain the correct answer
  - negative examples: (question, sentence) from the same paragraph which do not contain the answer

> *"What would a teacher do for someone who is cocky?"*

> *"The function of the teacher is to pressure the lazy, inspire the bored, deflate the cocky, encourage the timid, detect and correct individual flaws, and broaden the viewpoint of all."*

> *"How many people were lost in Algiers during 1620-21?"*

> *"Plague was present in at least one location in the Islamic world virtually every year between 1500 and 1850."*

MRPC

- Microsoft Research Paraphrase Corpus
- Sentence pairs automatically extracted from online news sources (with human annotations)
- To predict the sentences in the pair are semantically equivalent

> *"The decision to issue new guidance has been prompted by intelligence passed to Britain by the FBI in a secret briefing in late July ."*
>
> *"Scotland Yard 's decision to issue new guidance has been prompted by new intelligence passed to Britain by the FBI in late July ."*

> *"The company 's operating loss rose 59 percent to $ 73 million , from $ 46 million a year earlier ."*
>
> *"Operating revenue fell 4.5 percent to $ 2.3 billion from a year earlier ."*

Other GLUE tasks

- QQP (Quora Question Pairs)
  - binary classification task
  - 判斷兩個問句是否在語意上是相等的
- SST-2 (Stanford Sentiment Treebank)
  - binary single-sentence classification task
  - sentences extracted from movie reviews with human annotations of their sentiment
- CoLA (Corpus of Linguistic Acceptability)
  - a binary single-sentence classification task
  - predict whether an English sentence is linguistically “acceptable” or not

Other GLUE tasks

- STS-B (Semantic Textual Similarity Benchmark)
  - sentence pairs drawn from news headlines and other sources
  - annotated with a score from 1 to 5 denoting how similar the two sentences are in terms of semantic meaning
- RTE (Recognizing Textual Entailment)
  - binary entailment task (類似MNLI但是training data少很多)
- WNLI (Winograd NLI)
  - small natural language inference dataset

"I put the cake away in the refrigerator. It has a lot of butter in it."

"The cake has a lot of butter in it."

SQuAD v1.1

- Standford Question Answering Dataset
  - crowdsourced question/answer pair
- Given a question and a paragraph from Wikipedia containing the answer
  - predict the answer text span in the paragraph

- Input Question:

  `Where do water droplets collide with ice crystals to form precipitation?`

- Input Paragraph:

  `... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. ...`

- Output Answer:

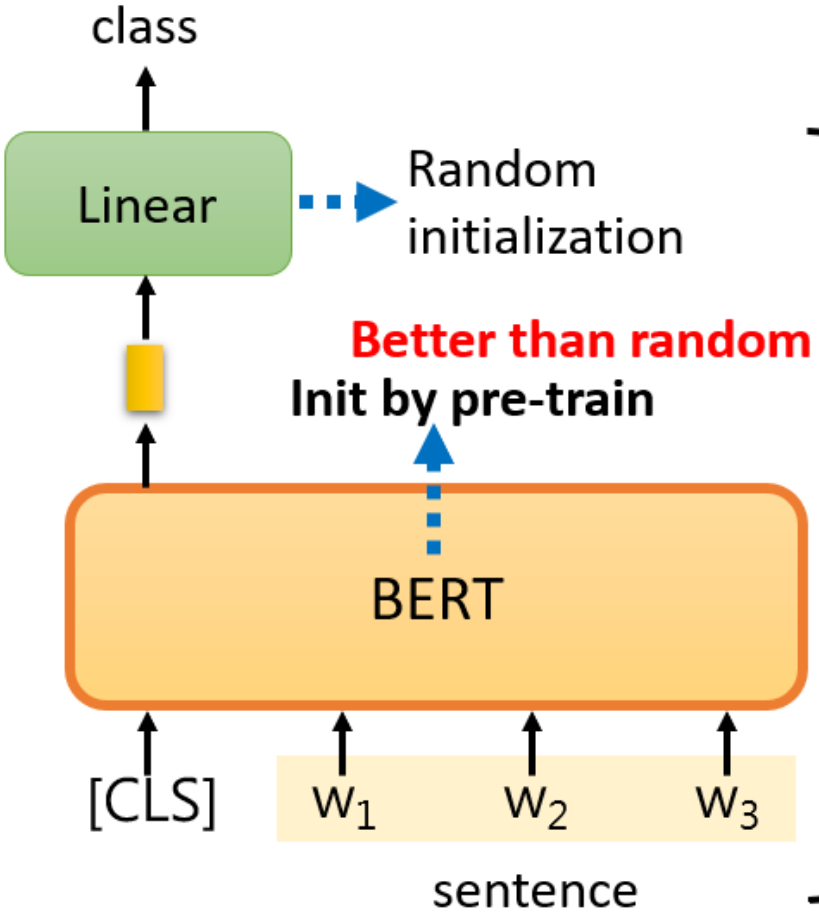
  `within a cloud`

CoNLL 2003 NER

- Named Entity Recognition (NER) dataset
- Label = Person, Organization, Location, Miscellaneous(混合), or Other (non-named entity)
- 為了可以跟WordPiece tokenizer相容，利用的是第一個sub-token來進行分類
  - 其他##開頭的subtoken不進行predict
  - 採用的是cased model (區分大小寫，其他task都是uncased)

```
Jim    Hen    ##son  was  a  puppet  ##eer
I-PER  I-PER  X      O    O  O       X
```

Case#1: Sentence Classification



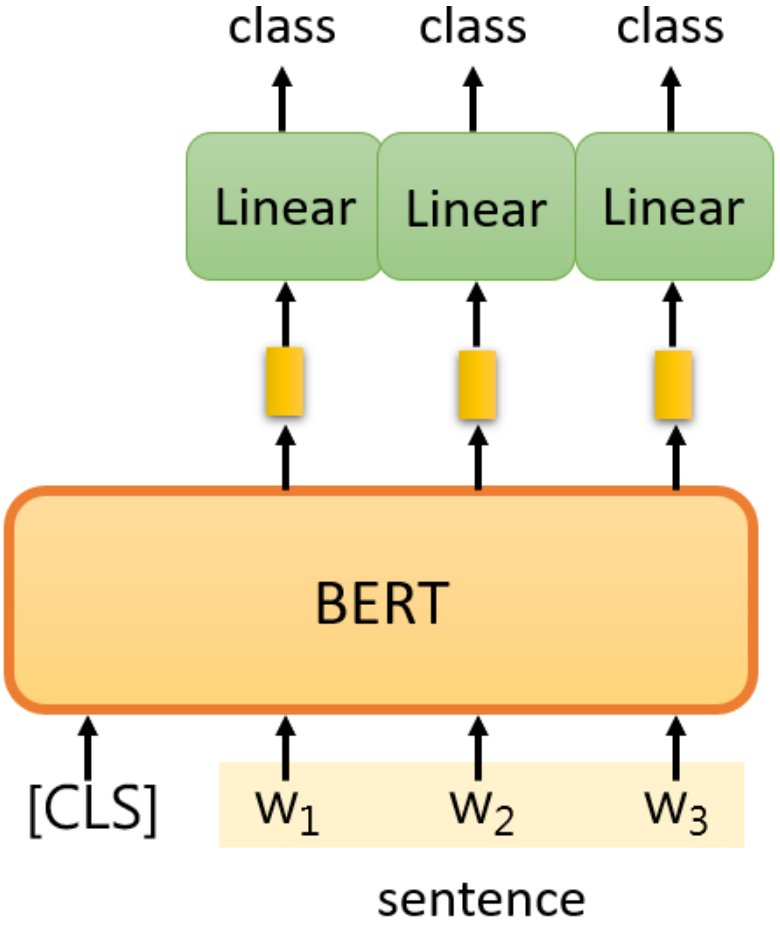
Input: sequence
output: class

Example:
Sentiment analysis

this is good
↓
positive

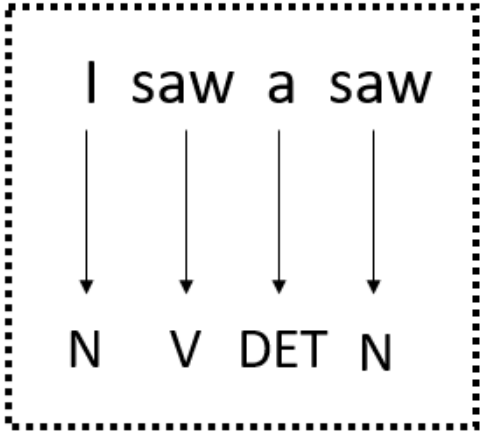
This is the model to be learned.

Case#2: Token Classification



Input: sequence
output: same as input

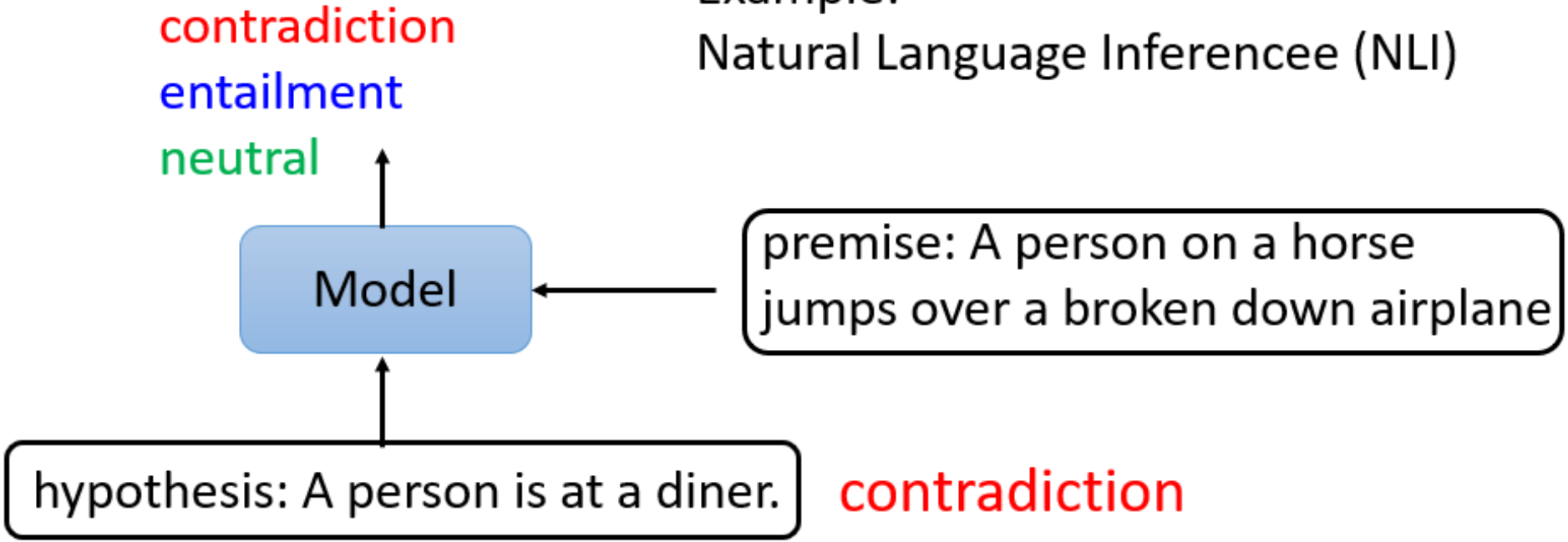
Example:
POS tagging



Case#3: Two-Sentence Relation

Input: two sequences
Output: a class

Example:
Natural Language Inferencee (NLI)



Case#4: Start/End Span Prediction (Extraction-based QA)

- Extraction-based Question Answering (QA)

Document: $D = \{d_1, d_2, \cdots, d_N\}$

Query: $Q = \{q_1, q_2, \cdots, q_M\}$

output: two integers (s, e)

Answer: $A = \{d_s, \cdots, d_e\}$

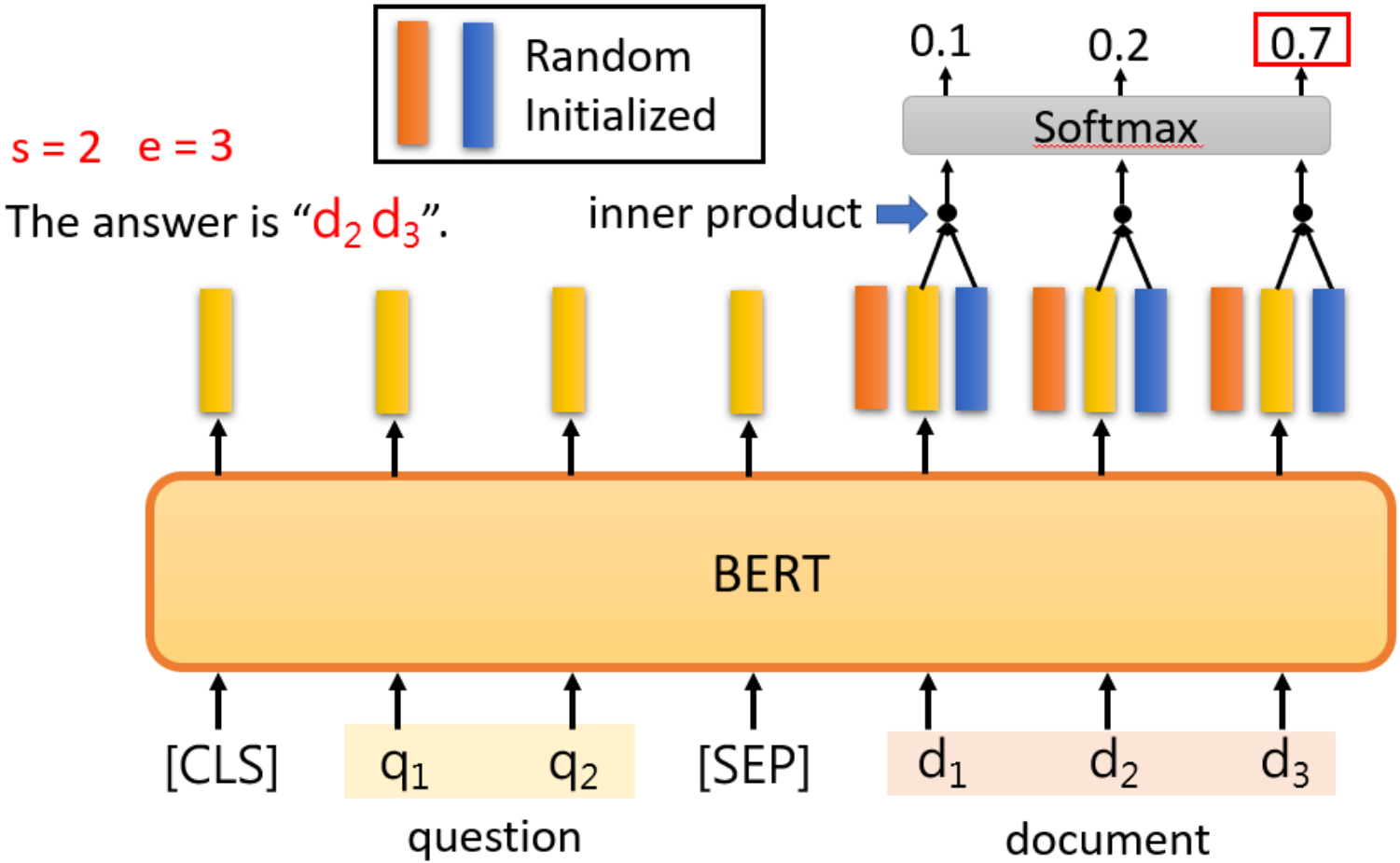
In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation includ…le, rain, sleet, snow, graupel and hail… 17 …pitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of …n scattered l… 77 …ns 79 are called "showers"

What causes precipitation to fall?
gravity $s = 17, e = 17$

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud $s = 77, e = 79$

Case#4: Start/End Span Prediction (Extraction-based QA)

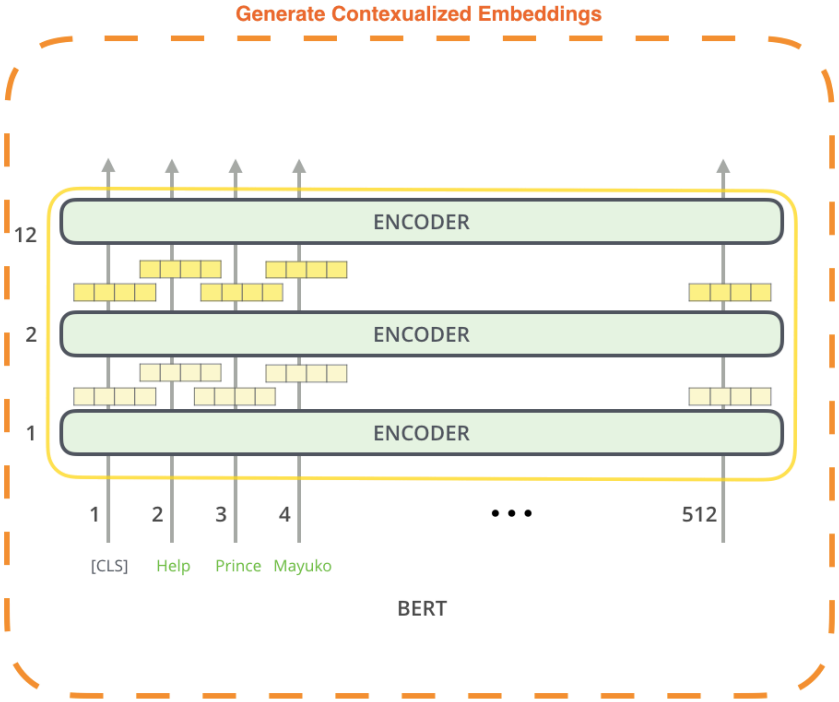


BERT for feature extraction

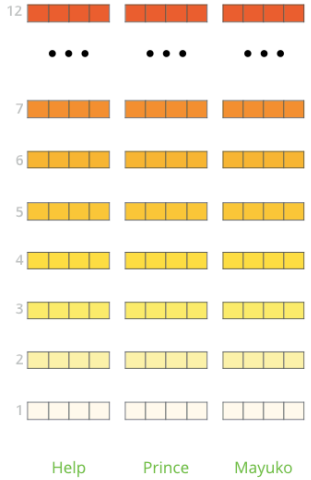
- Use the pre-trained BERT to create contextualized embeddings, and feed these embeddings to existing model優勢：
  - 作者認為並非所有的NLP task都可以表示成Transformer encoder input的樣子，因此可能會額外設計模型的架構
  - 只要先行運算過昂貴的pre-training 得到data的representation後，就可以依據這個表示做為基礎來訓練其他的模型，可以獲得很大的運算優勢

BERT for feature extraction

- Use pre-trained BERT to get contextualized embeddings and feed them into the task-specific models



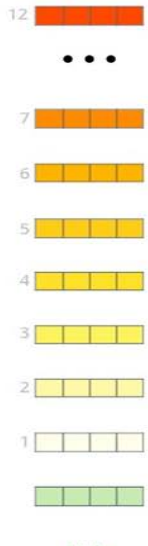
The output of each encoder layer along each token's path can be used as a feature representing that token.



But which one should we use?

BERT Embeddings Results on NER

What is the best contextualized embedding for "Help" in that context?
For named-entity recognition task CoNLL-2003 NER



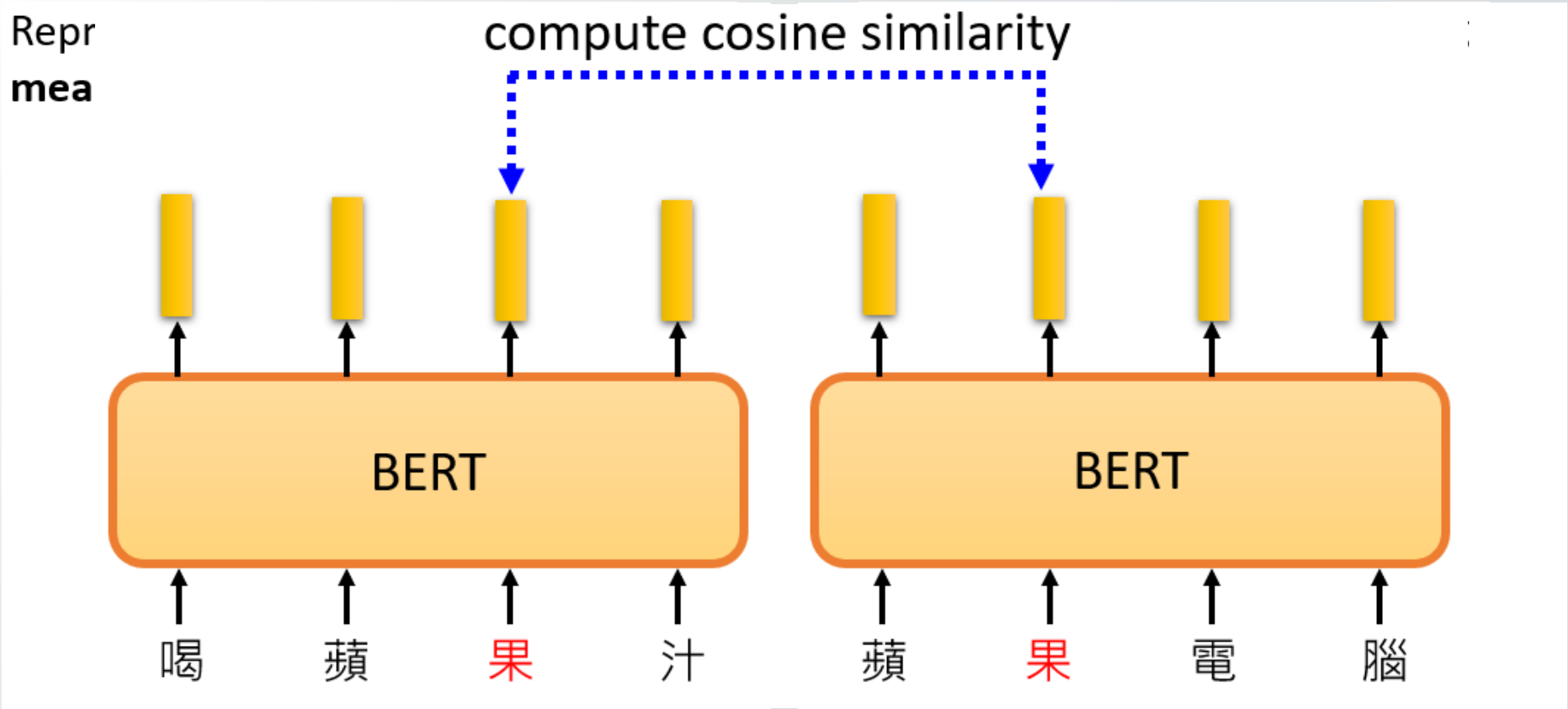
	Dev F1 Score
First Layer	91.0
Last Hidden Layer	94.9
Sum All 12 Layers	95.5
Second-to-Last Hidden Layer	95.6
Sum Last Four Hidden	95.9
Concat Last Four Hidden	96.1

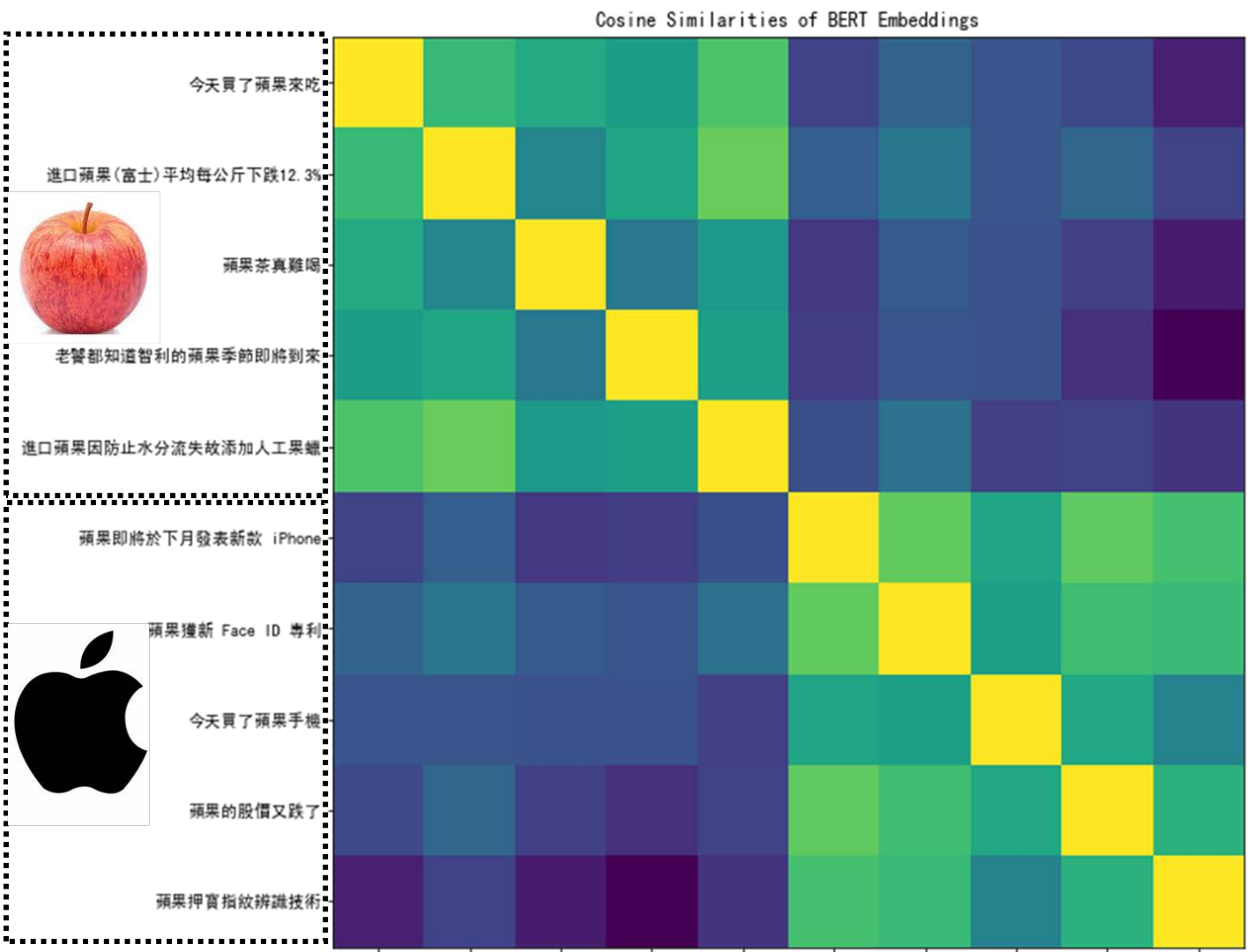
fine-tune = 96.4

Why does BERT work?

Pretrained model = Better?

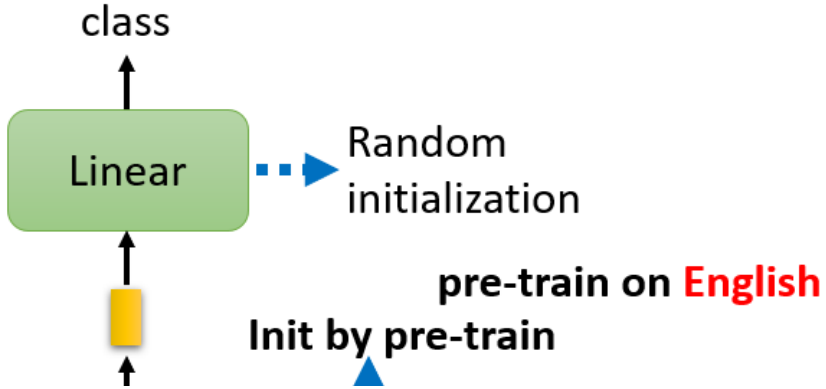
Contextualized Embeddings




Cosine Similarities of BERT Embeddings
今天買了蘋果來吃
進口蘋果(富士)平均每公斤下跌12.3%
蘋果茶真難喝
老饕都知道智利的蘋果季節即將到來
進口蘋果因防止水分流失故添加人工果蠟
蘋果即將於下月發表新款 iPhone
蘋果獲新 Face ID 專利
今天買了蘋果手機
蘋果的股價又跌了
蘋果押寶指紋辨識技術

DNA Sequence Classification

- Applying BERT to **protein, DNA, music classification**



	Protein			DNA				Music
	localization	stability	fluorescence	H3	H4	H3K9ac	Splice	composer
specific	69.0	76.0	63.0	87.3	87.3	79.1	94.1	-
BERT	64.8	74.5	63.7	83.0	86.2	78.3	97.5	55.2
re-emb	63.3	75.4	37.3	78.5	83.7	76.3	95.6	55.2
rand	58.6	65.8	27.5	75.6	66.5	72.8	95	36

DNA sequence ⋯▶ A G A C

BERT & its family

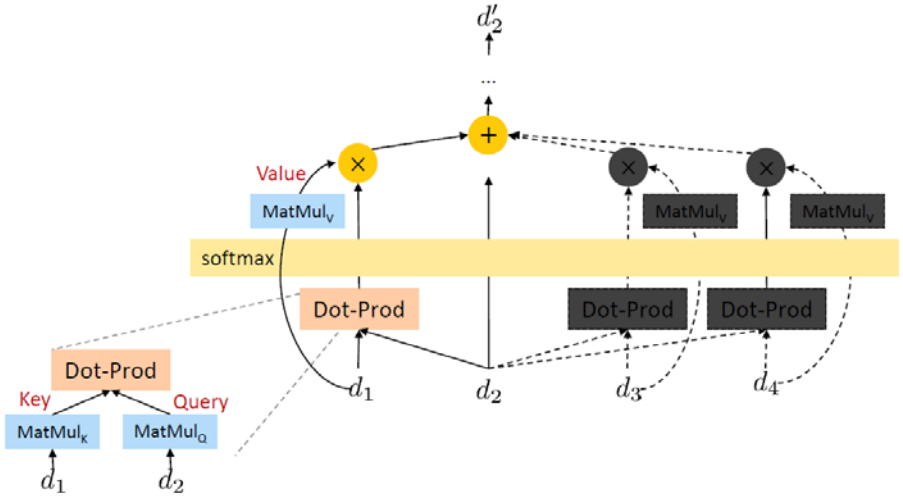
Various variations of BERT

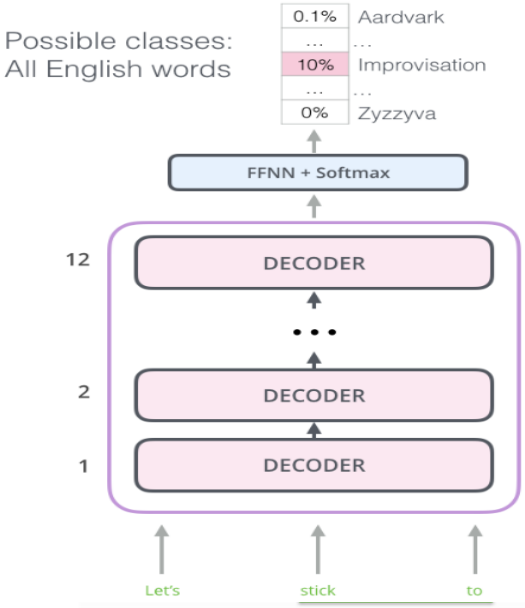
OpenAI GPT

- Left-to-right architecture (LTR Transformer decoder)
  - Constrained self-attention where every token can only attend to context to its left
  - Transformer decoder=>text generation
- Every token can only attended to previous tokens in the self-attention layers of the Transformer
- BERT作者認為當利用LTR的傳統模型架構來訓練token-level task時表現通常不太好
  - 例如: SQuAD question answering, 如果上下前後文都考慮其內容的話表現才會比較好

Decoder Self-Attention

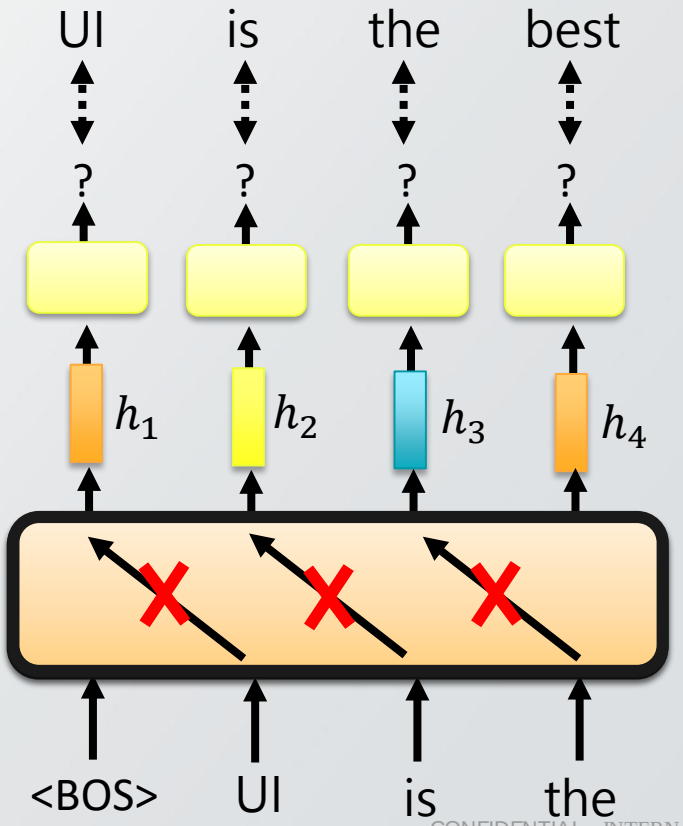
- In the decoder, the self-attention layer is only allowed to attend to earlier positions in the output sequence.
  - It's a natural choice for language modeling (predicting the next word) since it's built to mask future tokens – a valuable feature when it's generating a translation word by word
  - Masking future positions before the Softmax step

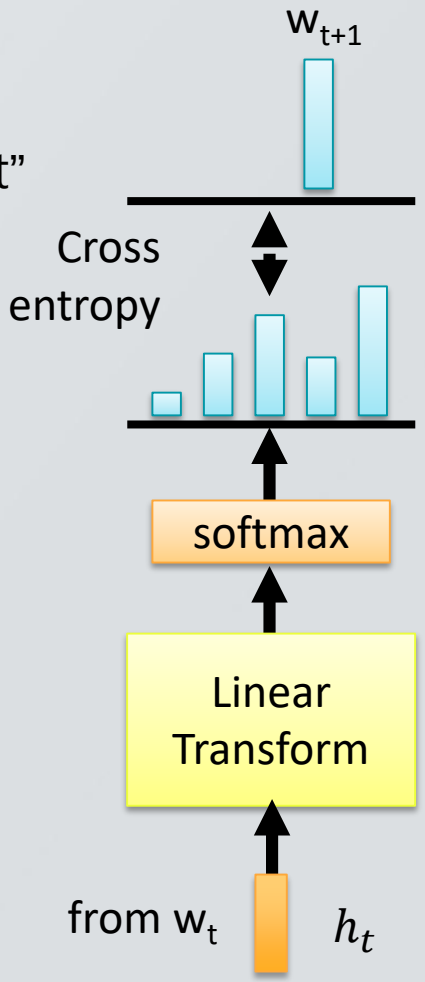




Pretraining task: Predict next token (CLM)

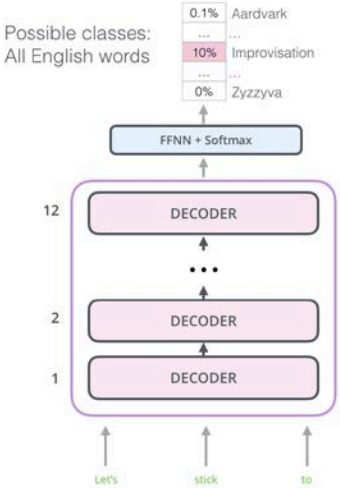
Training data:
"UI is the best"

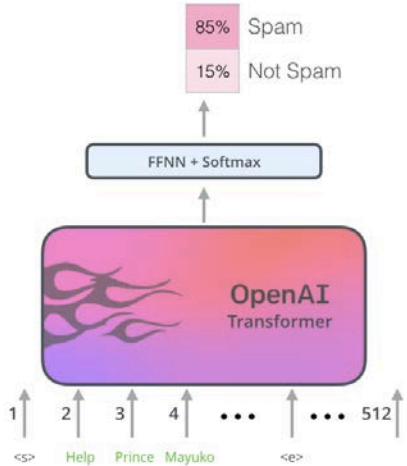




GPT Application #1: as BERT

- Train the model on the same language modeling task
  - Predict the next word using massive (unlabeled) datasets
- 當pre-trained完成以後，可以用至downstream task，如sentence classification (垃圾郵件分類器)



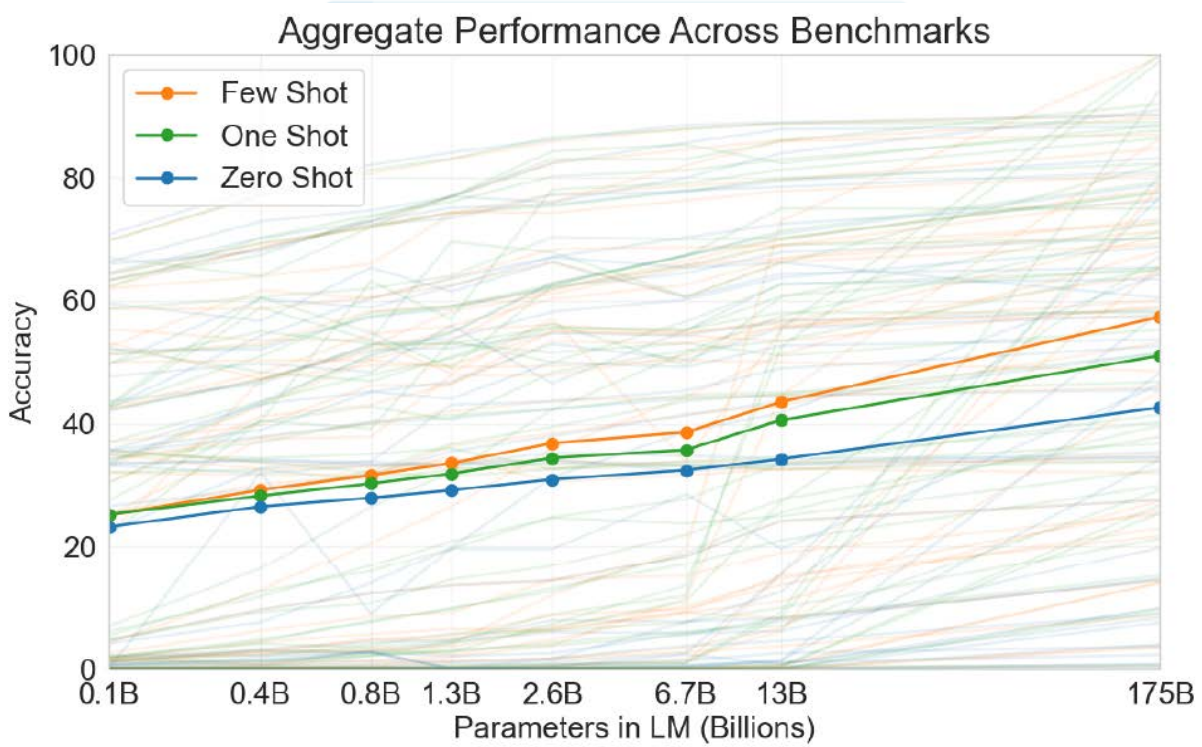


GPT Application #2: Text Generation

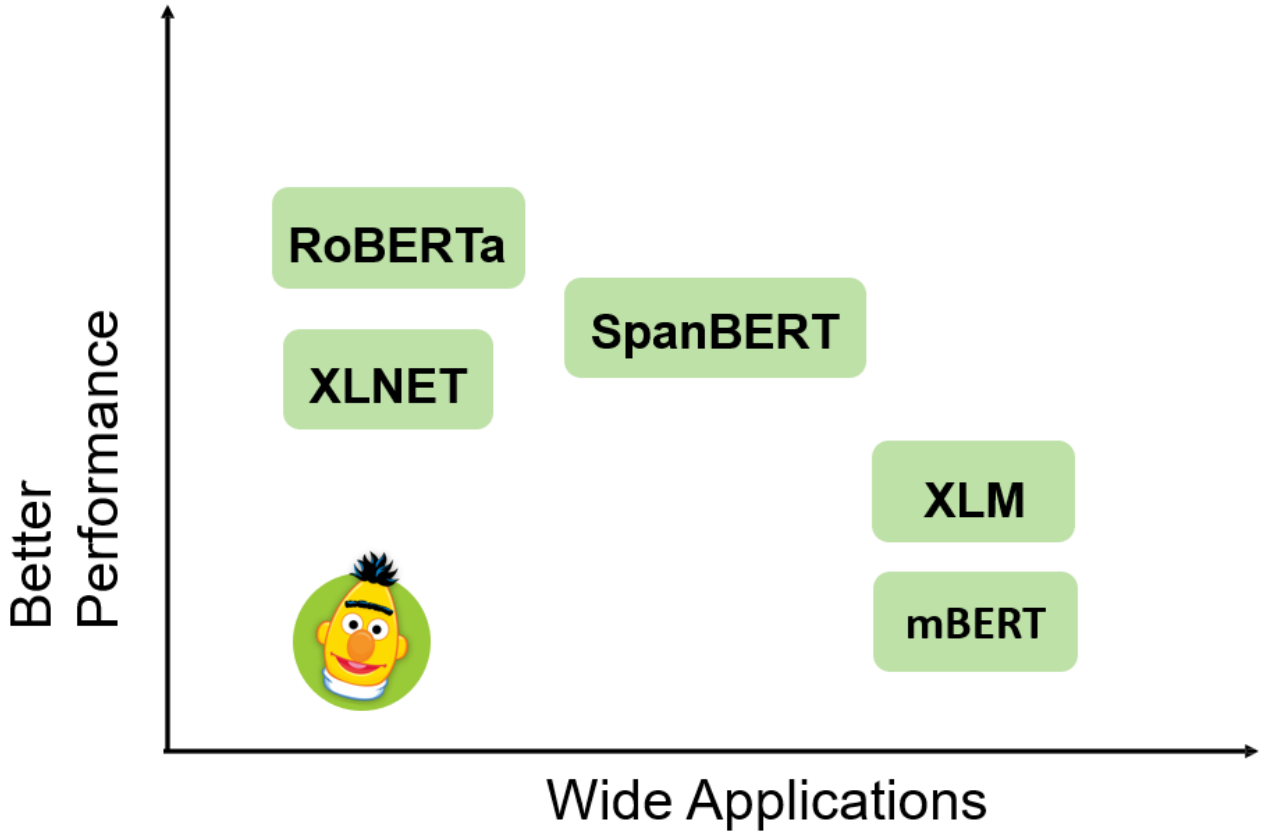
https://talktotransformer.com/

EM PROMPT -WRITTEN)	*In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*
MODEL MPLETION (MACHINE- 10 TRIES)	The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved. Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

GPT Application #3: "In-context" Learning



Beyond BERT



RoBERTa: Robustly Optimized BERT Pretraining Approach

1. Dynamic masking
   - Each sequence is masked in 10 different ways over the 40 epochs of training
     - Original masking is performed during data preprocessing
2. Optimization hyperparameters
   - Peak learning rate and number of warmup steps tuned separately for each setting
     - Training is very sensitive to the Adam epsilon term
     - Setting β2 = 0.98 improves stability when training with large batch sizes
3. Data
   A. Does not randomly inject short sequences
   B. Train only with full-length sequences
      - Original model trains with a reduced sequence length for first 90% of updates
   C. BookCorpus, CC-News, OpenWebText, Stories

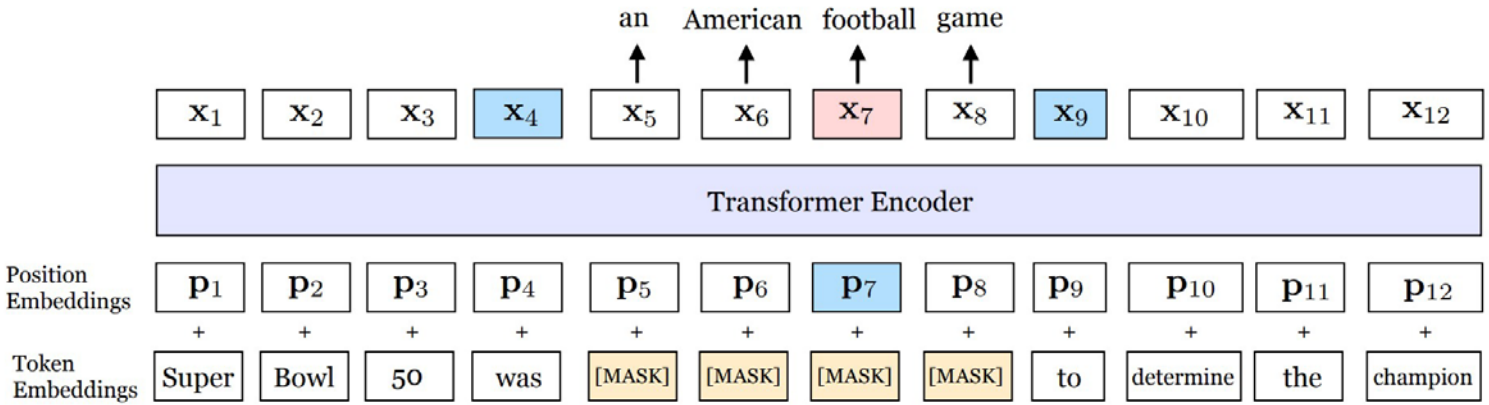
RoBERTa GLUE Results

	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS	WNLI	Avg
Single-task single models on dev										
$\text{BERT}_{\text{LARGE}}$	86.6/-	92.3	91.3	70.4	93.2	88.0	60.6	90.0	-	-
$\text{XLNet}_{\text{LARGE}}$	89.8/-	93.9	91.8	83.8	95.6	89.2	63.6	91.8	-	-
RoBERTa	**90.2/90.2**	**94.7**	**92.2**	**86.6**	**96.4**	**90.9**	**68.0**	**92.4**	**91.3**	-
Ensembles on test (from leaderboard as of July 25, 2019)										
ALICE	88.2/87.9	95.7	**90.7**	83.5	95.2	92.6	**68.6**	91.1	80.8	86.3
MT-DNN	87.9/87.4	96.0	89.9	86.3	96.5	92.7	68.4	91.1	89.0	87.6
XLNet	90.2/89.8	98.6	90.3	86.3	**96.8**	**93.0**	67.8	91.6	**90.4**	88.4
RoBERTa	**90.8/90.2**	**98.9**	90.2	**88.2**	96.7	92.3	67.8	**92.2**	89.0	**88.5**

SpanBERT: Improving Pre-training by Representing and Predicting Spans

1. Span masking
   - Random process to mask spans of tokens
2. Single sentence training
   - Single contiguous segment of text for each training sample (instead of two)
3. Span boundary objective (SBO)
   - Predict the entire masked span using only the span's boundary

$$\mathcal{L}(\text{football}) = \mathcal{L}_{\text{MLM}}(\mathbf{x}_7) + \mathcal{L}_{\text{SBO}}(\mathbf{x}_4, \mathbf{x}_9, \mathbf{p}_7)$$



SpanBERT Results

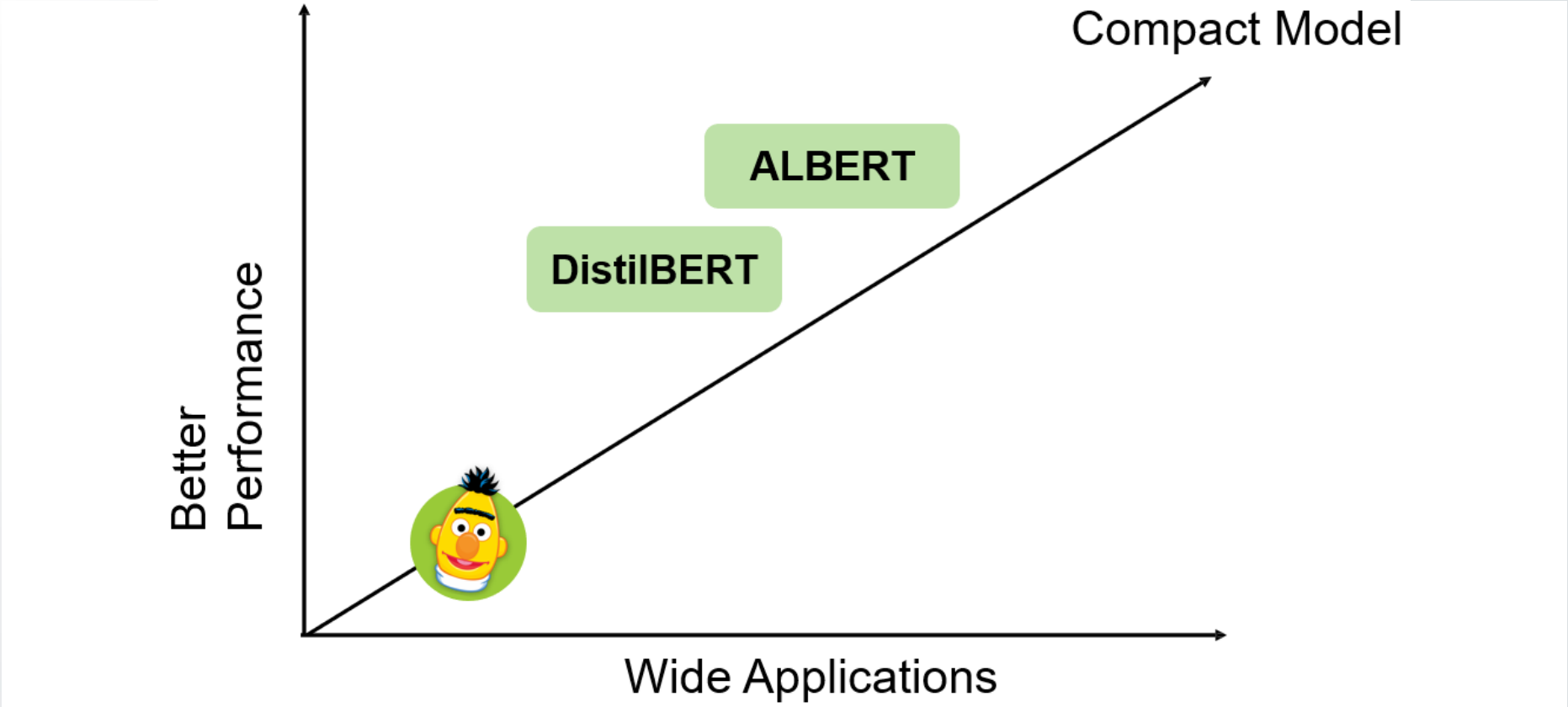
1. Masking scheme

	SQuAD 2.0	NewsQA	TriviaQA	Coreference	MNLI-m	QNLI
Subword Tokens	83.8	72.0	76.3	**77.7**	86.7	92.5
Whole Words	84.3	72.8	77.1	76.6	86.3	92.8
Named Entities	84.8	72.7	78.7	75.6	86.0	93.1
Noun Phrases	85.0	**73.0**	77.7	76.7	86.5	93.2
Random Spans	**85.4**	**73.0**	**78.8**	76.4	**87.0**	**93.3**

2. Auxiliary objective

	SQuAD 2.0	NewsQA	TriviaQA	Coreference	MNLI-m	QNLI
Span Masking (2seq) + NSP	85.4	73.0	78.8	76.4	87.0	93.3
Span Masking (1seq)	86.7	73.4	80.0	76.3	87.3	93.8
Span Masking (1seq) + SBO	**86.8**	**74.1**	**80.3**	**79.0**	**87.6**	**93.9**

Beyond BERT

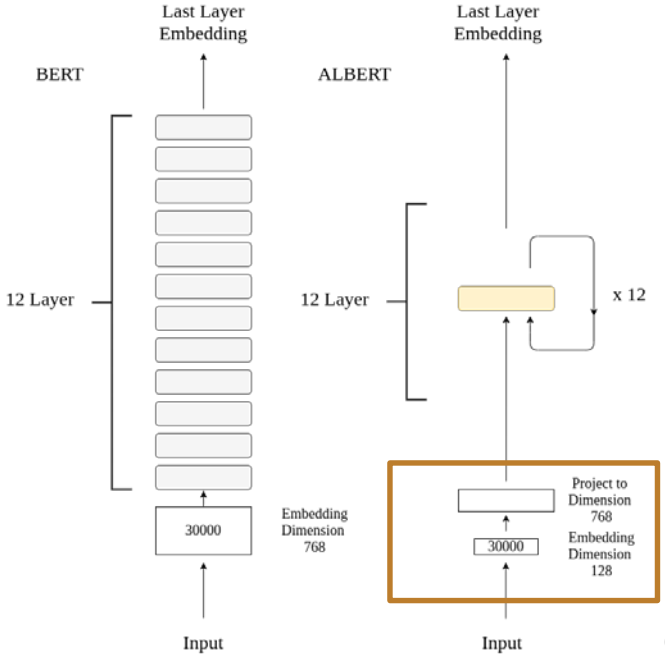


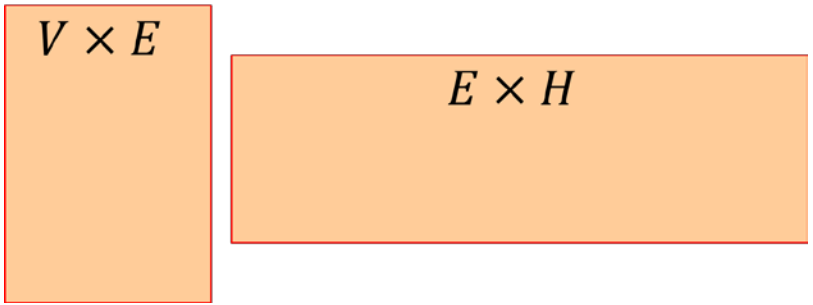
ALBERT: A Lite BERT

1. Factorized embedding parameterization

A. Original BERT WordPiece embedding size is tied with the hidden layer size (30000(V) x 768(H) = 23.04M Parameters)

B. ALBERT uses transformation/projection (30000(V) x128(E) = 3.8M, 128(E) x 768(H) = 0.098M Total = 3.898M Parameters)





ALBERT: A Lite BERT

Model	E	Parameters	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg
ALBERT base not-shared	64	87M	89.9/82.9	80.1/77.8	82.9	91.5	66.7	81.3
	128	89M	89.9/82.8	80.3/77.3	83.7	91.5	67.9	81.7
	256	93M	90.2/83.2	80.3/77.4	84.1	91.9	67.3	81.8
	768	108M	90.4/83.2	80.4/77.6	84.5	92.8	68.2	82.3
ALBERT base all-shared	64	10M	88.7/81.4	77.5/74.8	80.8	89.4	63.5	79.0
	128	12M	89.3/82.3	80.0/77.1	81.6	90.3	64.0	80.1
	256	16M	88.8/81.5	79.1/76.3	81.5	90.3	63.4	79.6
	768	31M	88.6/81.5	79.2/76.6	82.0	90.6	63.3	79.8

	Model	Parameters	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg
ALBERT base E=768	all-shared	31M	88.6/81.5	79.2/76.6	82.0	90.6	63.3	79.8
	shared-attention	83M	89.9/82.7	80.0/77.2	84.0	91.4	67.7	81.6
	shared-FFN	57M	89.2/82.1	78.2/75.4	81.5	90.8	62.6	79.5
	not-shared	108M	90.4/83.2	80.4/77.6	84.5	92.8	68.2	82.3
ALBERT base E=128	all-shared	12M	89.3/82.3	80.0/77.1	82.0	90.3	64.0	80.1
	shared-attention	64M	89.9/82.8	80.7/77.9	83.4	91.9	67.6	81.7
	shared-FFN	38M	88.9/81.6	78.6/75.6	82.3	91.7	64.4	80.2
	not-shared	89M	89.9/82.8	80.3/77.3	83.2	91.5	67.9	81.6

ALBERT Pre-training Task

3. Inter-sentence coherence loss

- Original BERT's NSP (next sentence prediction) contains both topical and ordering information
  - Topical cues help more → model utilizes more → model learns LM information less
- SOP (sentence order prediction) focuses on ordering not topical cues

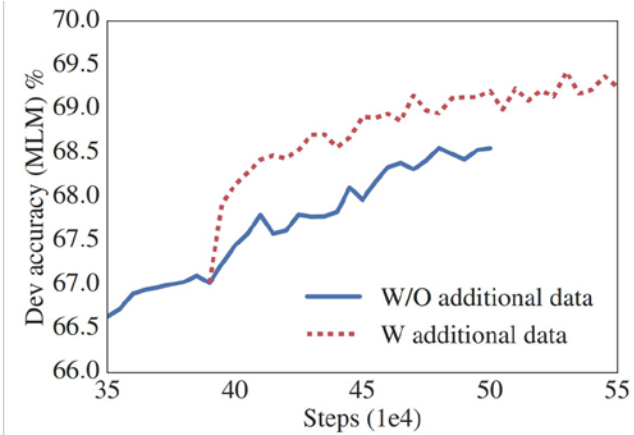
Mask Language Model:

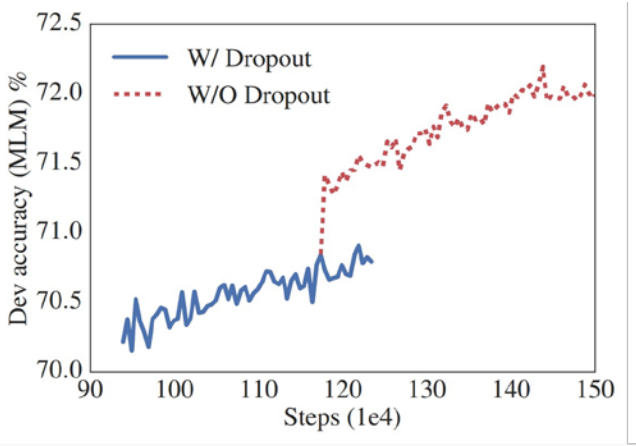
	Intrinsic Tasks			Downstream Tasks					
SP tasks	MLM	NSP	SOP	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg
None	54.9	52.4	53.3	88.6/81.5	78.1/75.3	81.5	89.9	61.7	79.0
NSP	54.5	90.5	52.0	88.4/81.5	77.2/74.6	81.6	**91.1**	62.3	79.2
SOP	54.0	78.9	86.5	**89.3/82.3**	**80.0/77.1**	**82.0**	90.3	**64.0**	**80.1**

Input: ✔ [CLS] 這隻手是人民的意志 [SEP] 人民的法槌

ALBERT Other Improvements

4. Additional data and removing dropout





	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg
No additional data	**89.3/82.3**	**80.0/77.1**	81.6	90.3	64.0	80.1
With additional data	88.8/81.7	79.1/76.3	**82.4**	**92.8**	**66.0**	**80.8**

	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg
With dropout	94.7/89.2	89.6/86.9	90.0	96.3	85.7	90.4
Without dropout	**94.8/89.5**	**89.9/87.2**	**90.4**	**96.5**	**86.1**	**90.7**

ALBERT GLUE Results

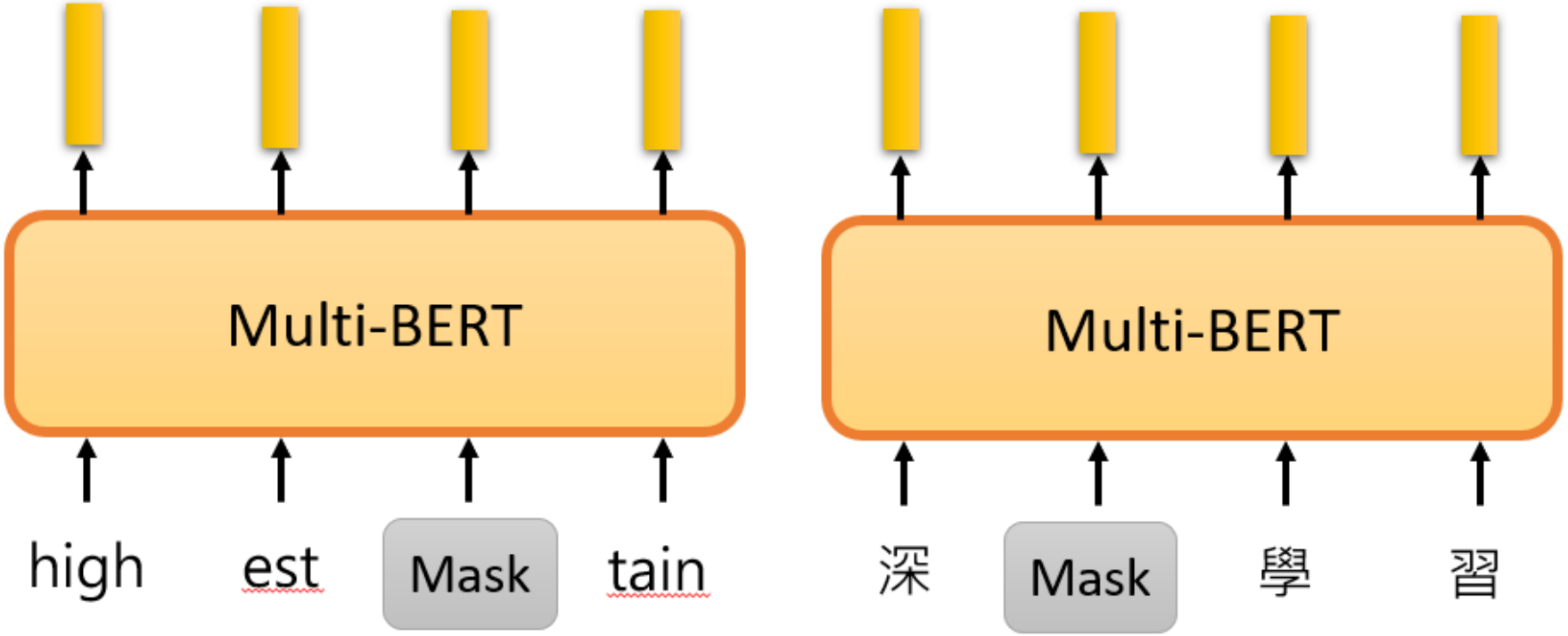
Models	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS	WNLI	Avg
Single-task single models on dev										
BERT-large	86.6	92.3	91.3	70.4	93.2	88.0	60.6	90.0	-	-
XLNet-large	89.8	93.9	91.8	83.8	95.6	89.2	63.6	91.8	-	-
RoBERTa-large	90.2	94.7	**92.2**	86.6	96.4	**90.9**	68.0	92.4	-	-
ALBERT (1M)	90.4	95.2	92.0	88.1	96.8	90.2	68.7	92.7	-	-
ALBERT (1.5M)	**90.8**	**95.3**	**92.2**	**89.2**	**96.9**	**90.9**	**71.4**	**93.0**	-	-
Ensembles on test (from leaderboard as of Sept. 16, 2019)										
ALICE	88.2	95.7	**90.7**	83.5	95.2	92.6	**69.2**	91.1	80.8	87.0
MT-DNN	87.9	96.0	89.9	86.3	96.5	92.7	68.4	91.1	89.0	87.6
XLNet	90.2	98.6	90.3	86.3	96.8	93.0	67.8	91.6	90.4	88.4
RoBERTa	90.8	98.9	90.2	88.2	96.7	92.3	67.8	92.2	89.0	88.5
Adv-RoBERTa	91.1	98.8	90.3	88.7	96.8	93.1	68.0	92.4	89.0	88.8
ALBERT	**91.3**	**99.2**	90.5	**89.2**	**97.1**	**93.4**	69.1	**92.5**	**91.8**	**89.4**

Multi-lingual BERT

Cross-lingual Language Models

mBERT: Pre-training

- Training a BERT model by many different languages.



mBERT: Experiments

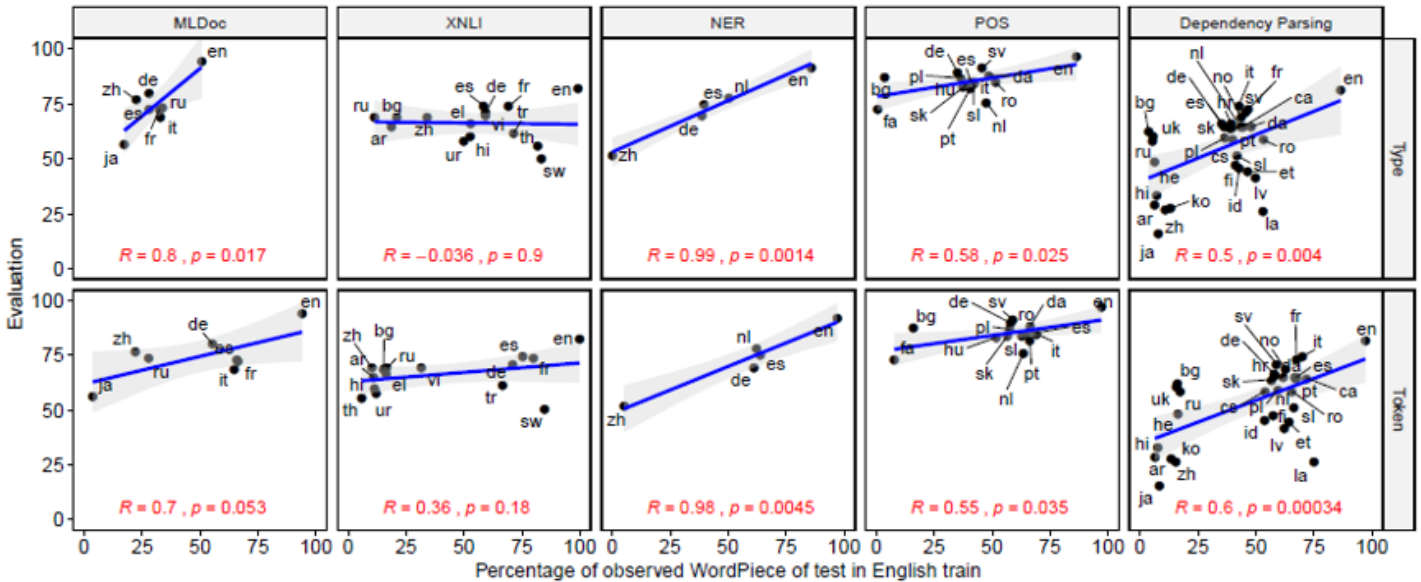
Fine-tuning \ Eval	EN	DE	NL	ES
EN	**90.70**	69.74	77.36	73.59
DE	73.83	**82.00**	76.25	70.03
NL	65.46	65.68	**89.86**	72.10
ES	65.38	59.40	64.39	**87.18**

Table 1: NER F1 results on the CoNLL data.

Fine-tuning \ Eval	EN	DE	ES	IT
EN	**96.82**	89.40	85.91	91.60
DE	83.99	**93.99**	86.32	88.39
ES	81.64	88.87	**96.71**	93.71
IT	86.79	87.82	91.28	**98.11**

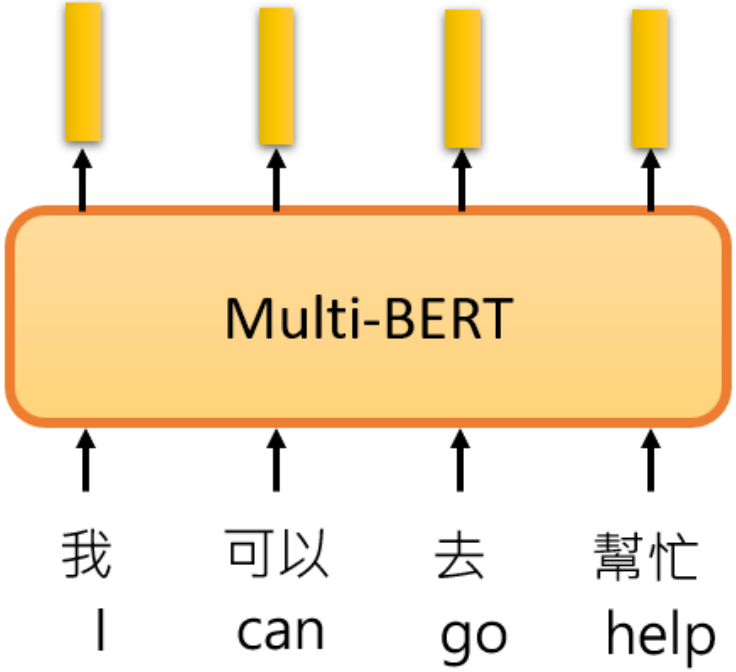
Table 2: POS accuracy on a subset of UD languages.

[Pires, et al., ACL'19]



[Wu, et al., EMNLP'19]

mBERT: Cross-lingual alignment



Mean Reciprocal Rank (MRR)

- Higher MRR, better alignment

Bi-lingual Dictionary

year → 年
month → 月
village → 村
big → 大

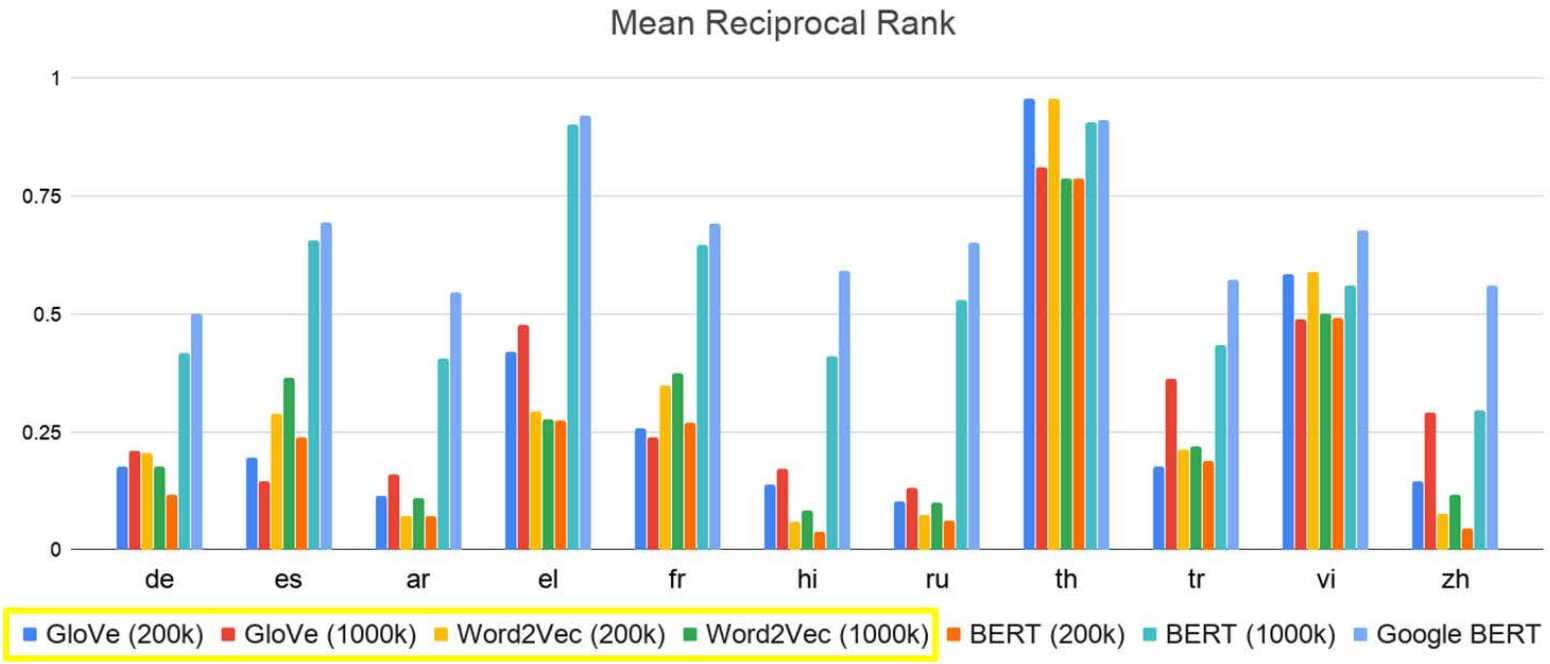
Cosine Similarity of Representation Vector

	年	月	和	村	人	大	他	....	Rank	Score
year	**0.7**	0.6	0.2	0.1	0.5	0.3	0.4		1	1/1
month	0.5	0.6	0.7	0.8	0.1	0.2	0.3		3	1/3
and	0.1	0.3	0.6	0.5	0.7	0.2	0.4		2	1/2
village	0.5	0.8	0.7	0.6	0.1	0.3	0.1		3	1/3
man	0.1	0.7	0.8	0.6	0.4	0.2	0.3		4	1/4
big	0.3	0.1	0.5	0.8	0.7	**0.9**	0.2		1	1/1
he	0.5	0.8	0.3	0.6	0.9	0.4	0.7		3	1/3
⋮										

Average

English to other language alignments results

1. The amount of training data is critical for alignment
2. Word2vec and GloVe cannot align well even with more data.



Why does mBERT will alignment?

Typical answers:

Different languages share some common tokens.

How do you explain Chinese v.s. English?

Code Switching

… DNA 的構造很像螺旋梯 …

(digits, punctuations)

Intermediate Language?

Language X shares tokens with Chinese and English.

Code Switching? Intermediate Language?

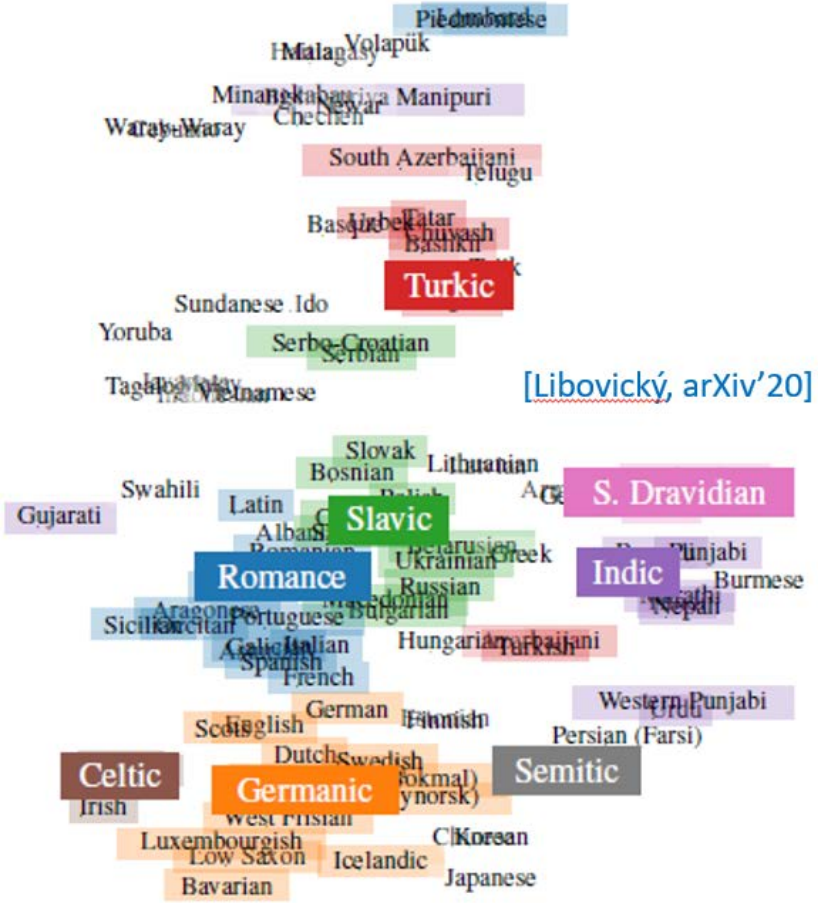
[K, et al., ICLR'20]

B-BERT	Train	Test	XNLI Accuracy	XNLI Wordpiece Contribution	NER Span F1-Score
en-es	en	es	72.3	1.4	61.9 (±0.8)
enfake-es	enfake		**70.9**		**62.6 (±1.6)**
en-hi	en	hi	60.1	0.5	61.6 (±0.7)
enfake-hi	enfake		**59.6**		**62.9 (±0.7)**
en-ru	en	ru	66.4	0.7	57.1* (±0.9)
enfake-ru	enfake		**65.7**		**54.2 (±0.7)**
en-enfake	enfake	enfake	78.0	0.5	78.9* (±0.7)
en-enfake	enfake	en	**77.5**		76.6(±0.8)

English: the cat is a good cat

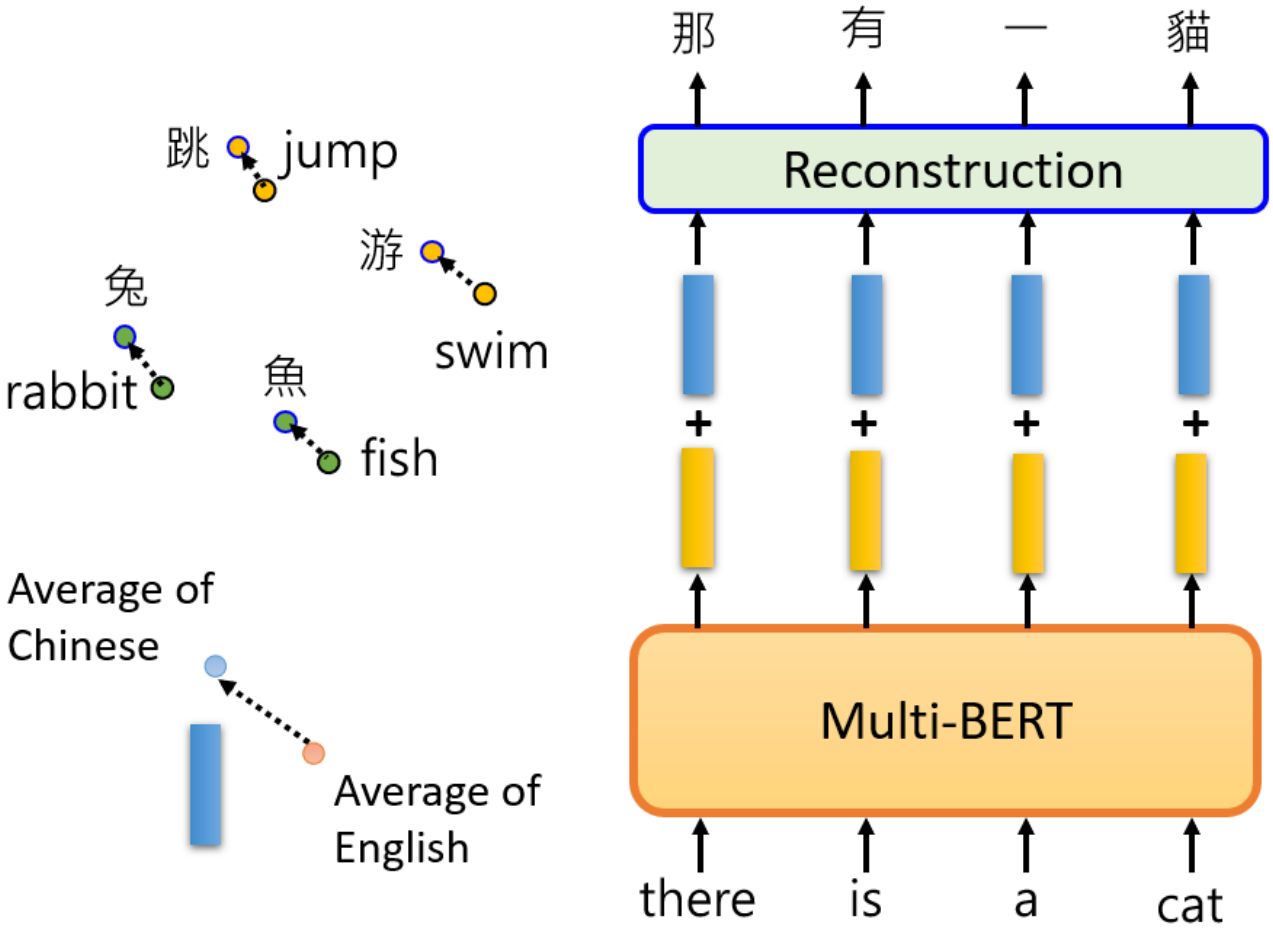
Fake-English: 甲 乙 天 地 人 乙

Language clustering by mBERT

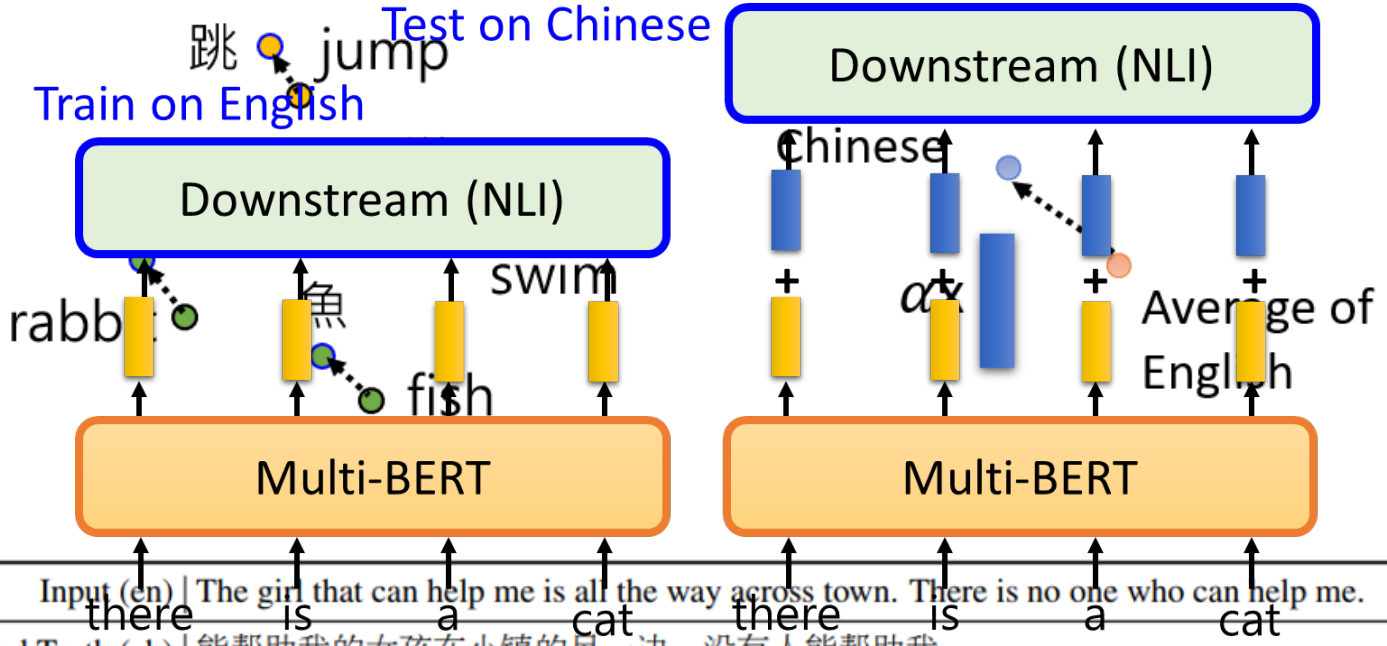


[Libovický, arXiv'20]

Implicitly language information in mBERT



Language information experiments



Input (en) | The girl that can help me is all the way across town. There is no one who can help me.

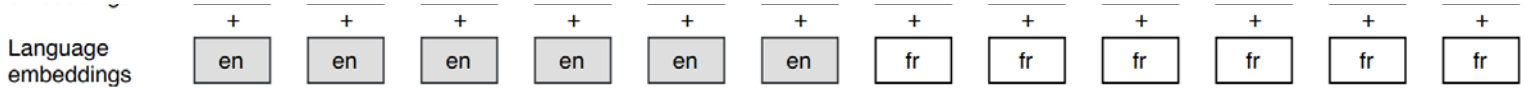
Ground Truth (zh) | 能帮助我的女孩在小镇的另一边。没有人能帮助我。。

	en	de	es	ar	el	fr	hi	ru	th	tr	vi	zh	avg w/o en
finetune all layers	80.86	67.05	70.6	59.94	56.67	70.42	46.65	66.97	41.70	50.10	40.68	67.20	58.00
finetune last 6 layers	82.16	69.34	74.73	63.91	61.52	72.95	51.00	69.46	48.94	57.23	40.76	69.96	61.80
+ MDS	-	**69.72**	74.57	**64.53**	**62.02**	72.42	**51.28**	**69.56**	**49.90**	**57.45**	**43.07**	**70.16**	**62.24**
shifting weight		1.5	0.2	1.0	0.9	0.7	0.3	1.0	0.4	0.4	0.5	2.0	

XLM: Cross-lingual Language Model Pretraining

- Masked LM + Translation LM

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Δ
Machine translation baselines (TRANSLATE-TRAIN)																
Devlin et al. (2018)	81.9	-	77.8	75.9	-	-	-	-	70.7	-	-	76.6	-	-	61.6	-
XLM (MLM+TLM)	85.0	80.2	80.8	80.3	78.1	79.3	78.1	74.7	76.5	76.6	75.5	78.6	72.3	70.9	63.2	76.7
Machine translation baselines (TRANSLATE-TEST)																
Devlin et al. (2018)	81.4	-	74.9	74.4	-	-	-	-	70.4	-	-	70.1	-	-	62.1	-
XLM (MLM+TLM)	85.0	79.0	79.5	78.1	77.8	77.6	75.5	73.7	73.7	70.8	70.4	73.6	69.0	64.7	65.1	74.2
Evaluation of cross-lingual sentence encoders																
Conneau et al. (2018b)	73.7	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4	65.6
Devlin et al. (2018)	81.4	-	74.3	70.5	-	-	-	-	62.1	-	-	63.8	-	-	58.3	-
Artetxe and Schwenk (2018)	73.9	71.9	72.9	72.6	73.1	74.2	71.5	69.7	71.4	72.0	69.2	71.4	65.5	62.2	61.0	70.2
XLM (MLM)	83.2	76.5	76.3	74.2	73.1	74.0	73.1	67.8	68.5	71.2	69.2	71.9	65.7	64.6	63.4	71.5
XLM (MLM+TLM)	**85.0**	**78.7**	**78.9**	**77.8**	**76.6**	**77.4**	**75.3**	**72.5**	**73.1**	**76.1**	**73.2**	**76.5**	**69.6**	**68.4**	**67.3**	**75.1**



Reference

1. [Lewis, et al., arXiv'19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer, BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, arXiv, 2019
2. [Raffel, et al., arXiv'19] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, arXiv, 2019
3. [Joshi, et al., TACL'20] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, Omer Levy, SpanBERT: Improving Pre-training by Representing and Predicting Spans, TACL, 2020
4. [Song, et al., ICML'19] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, Tie-Yan Liu, MASS: Masked Sequence to Sequence Pre-training for Language Generation, ICML, 2019
5. [Zafrir, et al., NeurIPS workshop 2019] Ofir Zafrir, Guy Boudoukh, Peter Izsak, Moshe Wasserblat, Q8BERT: Quantized 8Bit BERT, NeurIPS workshop 2019

Reference

1. [Houlsby, et al., ICML'19] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, Sylvain Gelly, Parameter-Efficient Transfer Learning for NLP, ICML, 2019
2. [Hao, et al., EMNLP'19] Yaru Hao, Li Dong, Furu Wei, Ke Xu, Visualizing and Understanding the Effectiveness of BERT, EMNLP, 2019
3. [Liu, et al., arXiv'19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv, 2019
4. [Sanh, et al., NeurIPS workshop's] Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, NeurIPS workshop, 2019
5. [Jian, et al., arXiv'19] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, Qun Liu, TinyBERT: Distilling BERT for Natural Language Understanding, arXiv, 19

Reference

1. [Shoeybi, et al., arXiv'19]Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, Bryan Catanzaro, Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism, arXiv, 19
2. [Lan, et al., ICLR'20]Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut, ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, ICLR, 2020
3. [Kitaev, et al., ICLR'20] Nikita Kitaev, Lukasz Kaiser, Anselm Levskaya, Reformer: The Efficient Transformer, ICLR, 2020
4. [Beltagy, et al., arXiv'20] Iz Beltagy, Matthew E. Peters, Arman Cohan, Longformer: The Long-Document Transformer, arXiv, 2020
5. [Dai, et al., ACL'19] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, Ruslan Salakhutdinov, Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context, ACL, 2019
6. [Peters, et al., NAACL'18] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer, Deep contextualized word representations, NAACL, 2018

Reference

1. [Sanh, et al., NeurIPS workshop's] Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, NeurIPS workshop, 2019
2. [Jian, et al., arXiv'19] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, Qun Liu, TinyBERT: Distilling BERT for Natural Language Understanding, arXiv, 19
3. [Sun, et al., ACL'20] Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, Denny Zhou, MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices, ACL, 2020
4. [Zafrir, et al., NeurIPS workshop 2019] Ofir Zafrir, Guy Boudoukh, Peter Izsak, Moshe Wasserblat, Q8BERT: Quantized 8Bit BERT, NeurIPS workshop 2019
5. [Sun, et al., ACL'20] Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, Denny Zhou, MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices, ACL, 2020

Reference

1. [Pennington, et al., EMNLP'14] Jeffrey Pennington, Richard Socher, Christopher Manning, Glove: Global Vectors for Word Representation, EMNLP, 2014
2. [Mikolov, et al., NIPS'13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, Jeff Dean, Distributed Representations of Words and Phrases and their Compositionality, NIPS, 2013
3. [Bojanowski, et al., TACL'17] Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov, Enriching Word Vectors with Subword Information, TACL, 2017
4. [Su, et al., EMNLP'17] Tzu-Ray Su, Hung-Yi Lee, Learning Chinese Word Representations From Glyphs Of Characters, EMNLP, 2017
5. [Liu, et al., ACL'19] Xiaodong Liu, Pengcheng He, Weizhu Chen, Jianfeng Gao, Multi-Task Deep Neural Networks for Natural Language Understanding, ACL, 2019
6. [Stickland, et al., ICML'19] Asa Cooper Stickland, Iain Murray, BERT and PALs: Projected Attention Layers for Efficient Adaptation in Multi-Task Learning, ICML, 2019

Reference

1. [Howard, et al., ACL'18] Jeremy Howard, Sebastian Ruder, Universal Language Model Fine-tuning for Text Classification, ACL, 2018
2. [Alec, et al., 2018] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, Improving Language Understanding by Generative Pre-Training, 2018
3. [Devlin, et al., NAACL'19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL, 2019
4. [Alec, et al., 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, Language Models are Unsupervised Multitask Learners, 2019
5. [Want, et al., ICLR'20] Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, Luo Si, StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding, ICLR, 2020
6. [Yang, et al., NeurIPS'19] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le, XLNet: Generalized Autoregressive Pretraining for Language Understanding, NeurIPS, 2019

Reference

1. [Cui, et al., arXiv'19] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, Guoping Hu, Pre-Training with Whole Word Masking for Chinese BERT, arXiv, 2019
2. [Sun, et al., ACL'19] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, Hua Wu, ERNIE: Enhanced Representation through Knowledge Integration, ACL, 2019
3. [Dong, et al., NeurIPS'19] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, Hsiao-Wuen Hon, Unified Language Model Pre-training for Natural Language Understanding and Generation, NeurIPS, 2019

Reference

1. [K, et al., ICLR'20] Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual BERT: An empirical study, ICLR, 2020
2. [Pires, et al., ACL'19] Telmo Pires, Eva Schlinger, Dan Garrette, How multilingual is Multilingual BERT?, ACL, 2019
3. [Wu, et al., EMNLP'19] Shijie Wu, Mark Dredze, Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT, EMNLP, 2019
4. [Hsu, Liu, et al., EMNLP'19] Tsung-Yuan Hsu, Chi-Liang Liu and Hung-yi Lee, "Zero-shot Reading Comprehension by Cross-lingual Transfer Learning with Multi-lingual Language Representation Model", EMNLP, 2019
5. [Liu, et al., arXiv'20] Chi-Liang Liu, Tsung-Yuan Hsu, Yung-Sung Chuang, Hung-Yi Lee, A Study of Cross-Lingual Ability and Language-specific Information in Multilingual BERT, arXiv, 2020

Reference

1. [Hu, et al., arXiv'20] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, Melvin Johnson, XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization, arXiv, 2020
2. [Libovický, arXiv'20] Jindřich Libovický, Rudolf Rosa, Alexander Fraser, On the Language Neutrality of Pre-trained Multilingual Representations, arXiv, 2020

That's all!

Slido:
#BERTology_Leo
Feedback:
https://forms.gle/YGcdxwjcma6LP3iK6


UI.COM