1-of-N Encoding

apple = [1 0 0 0 0]

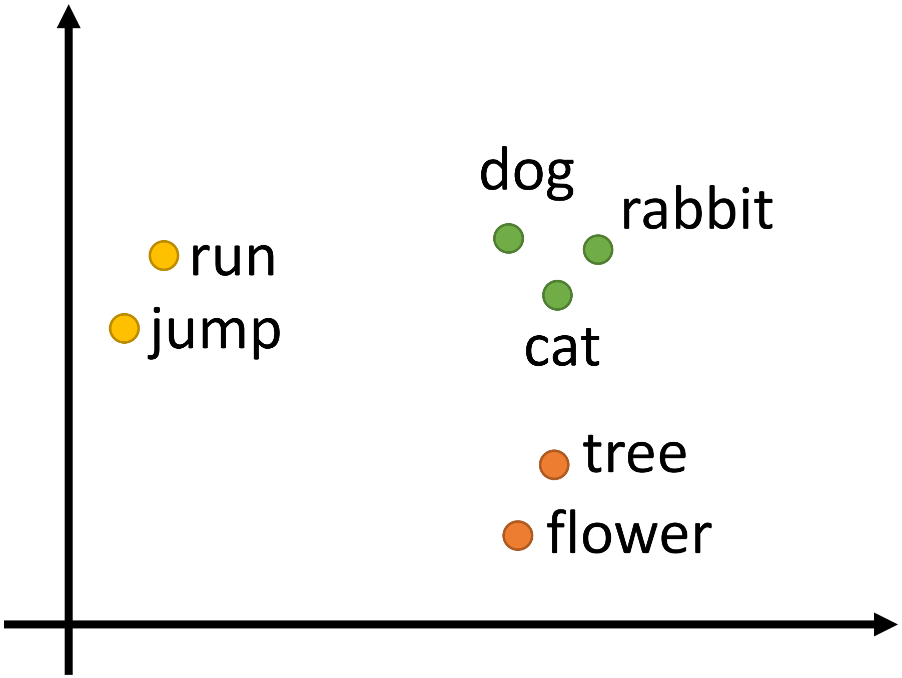
bag = [0 1 0 0 0]

cat = [0 0 1 0 0]

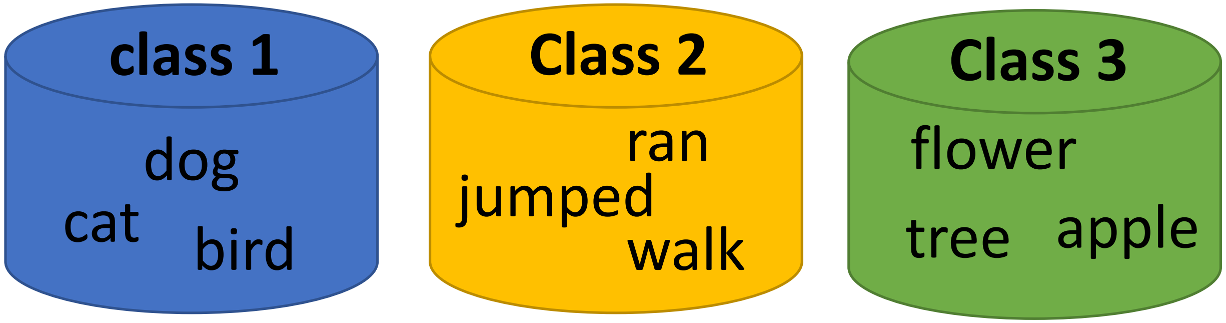
dog = [0 0 0 1 0]

elephant = [0 0 0 0 1]

Word Embedding



Word Class



class 1

dog
cat
bird

Class 2

ran
jumped
walk

Class 3

flower
tree
apple

https://arxiv.org/abs/1902.06006

A word can have multiple senses.

Have you paid that money to the bank yet ?

It is safest to deposit your money in the bank .

The victim was found lying dead on the river bank .

They stood on the river bank to fish.

The hospital has its own blood bank.

The third sense or not?

More Examples

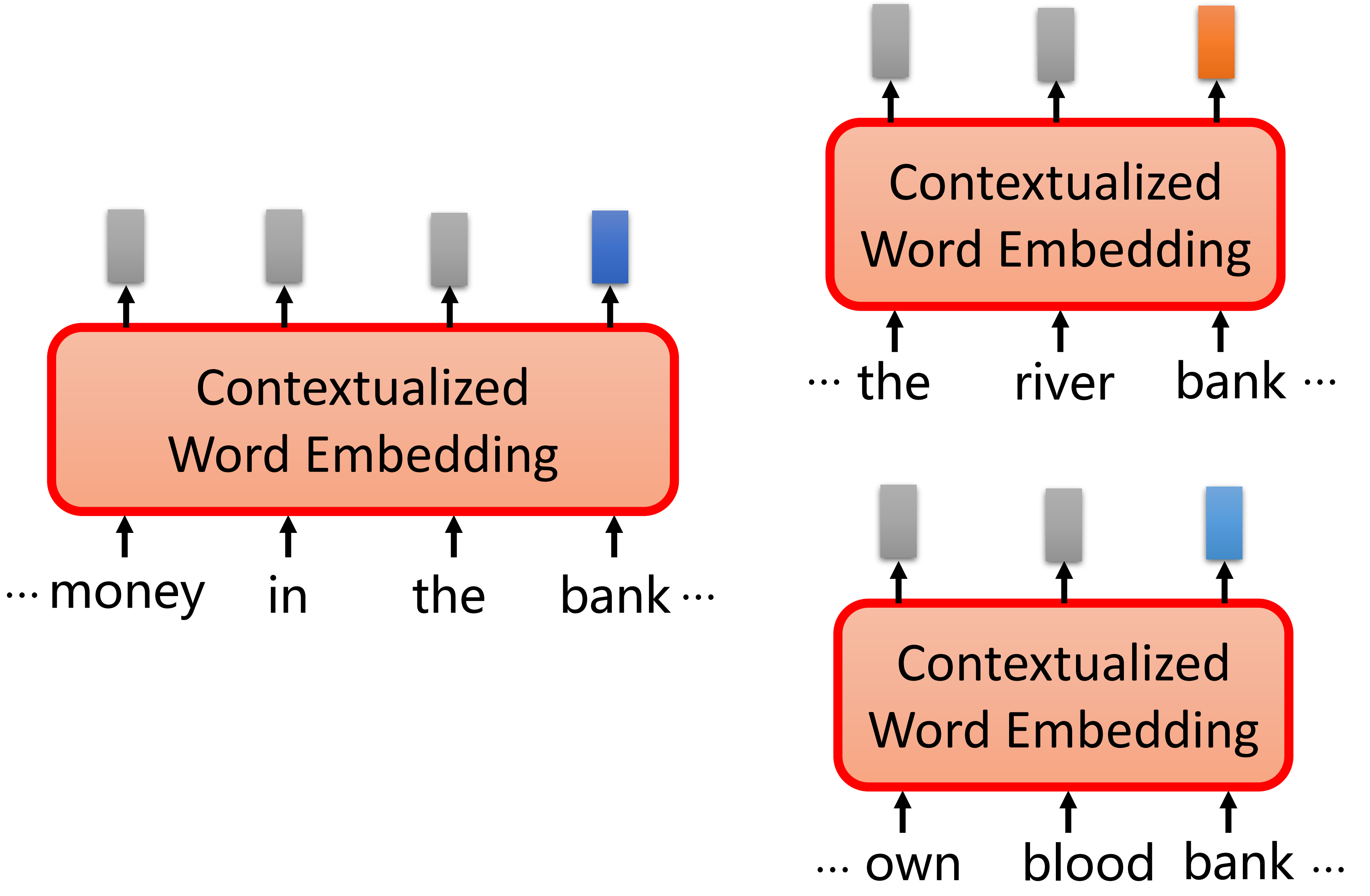
這是
加賀號護衛艦

他是尼祿

她也是尼祿

這也是加賀
號護衛艦

Contextualized Word Embedding

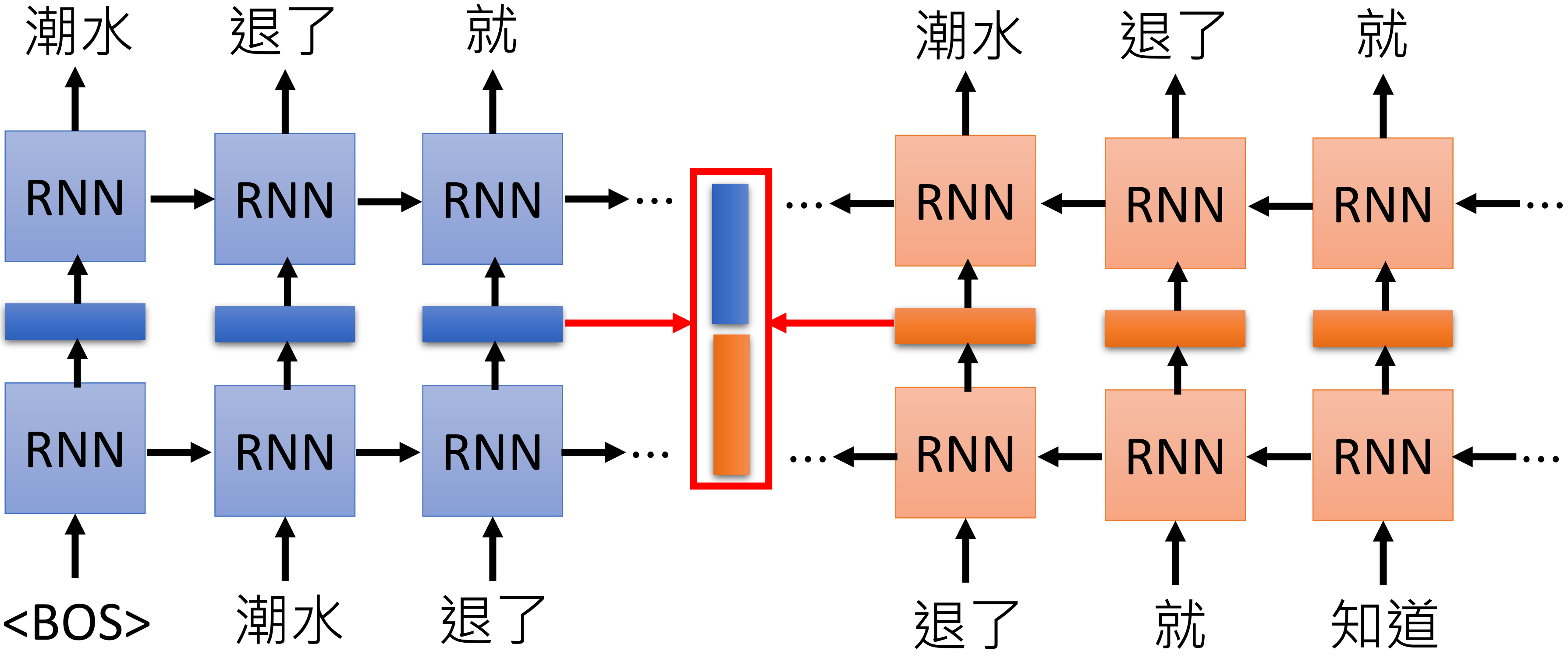


Embeddings from Language Model (ELMO)

https://arxiv.org/abs/1802.05365

- RNN-based language models (trained from lots of sentences)

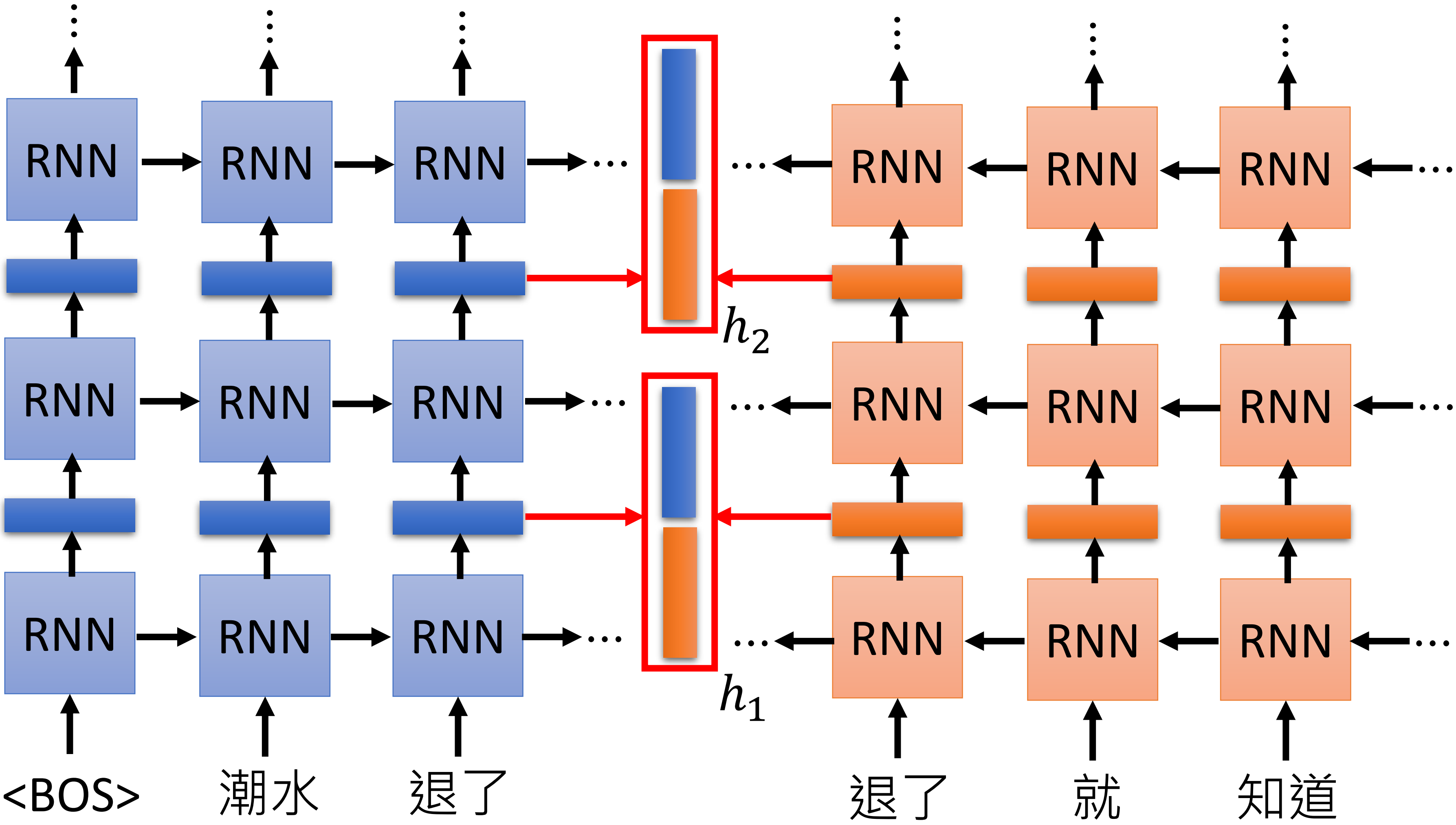
e.g. given “潮水 退了 就 知道 誰 沒穿 褲子”



ELMO

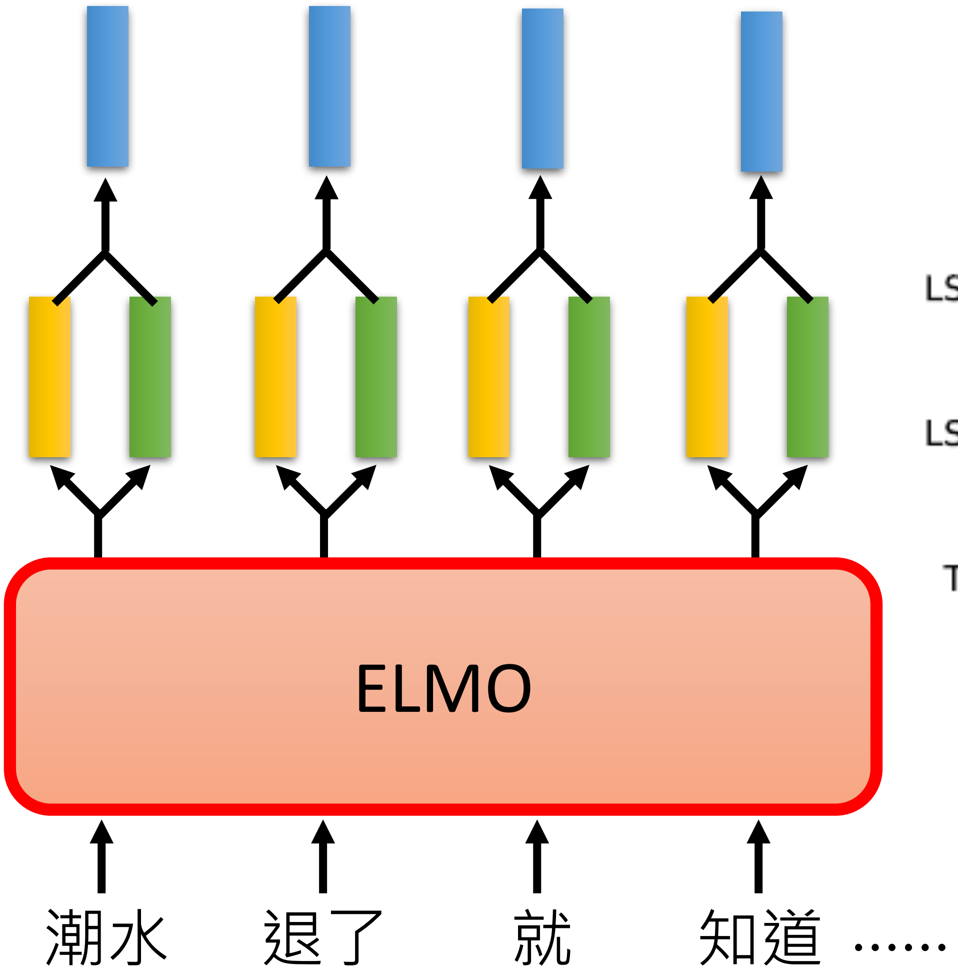
Each layer in deep LSTM can generate a latent representation.

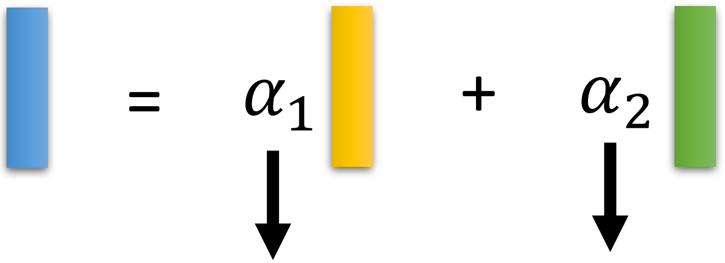
Which one should we use???



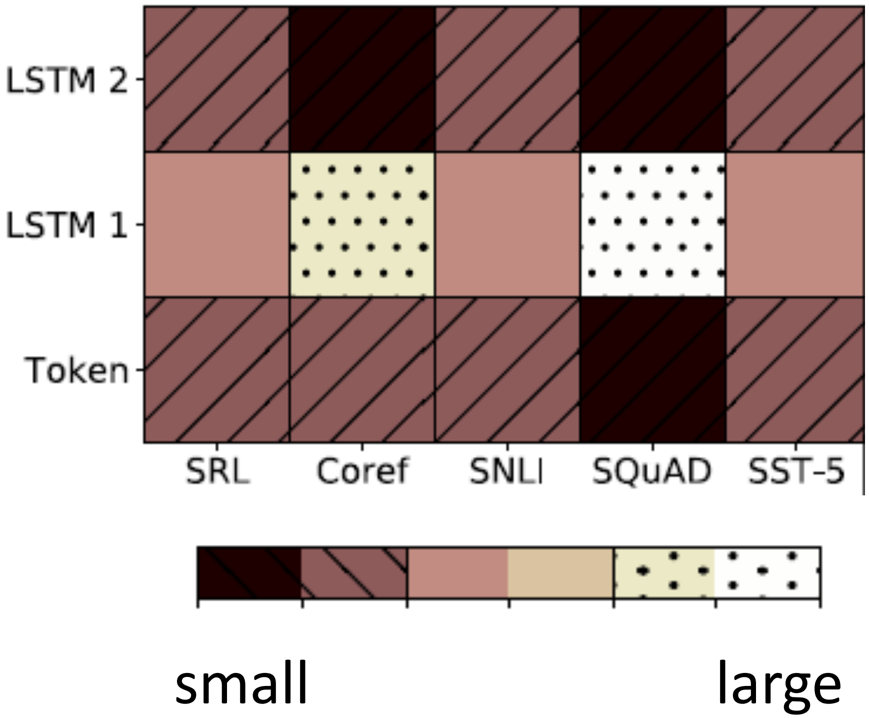
我全都要

ELMO





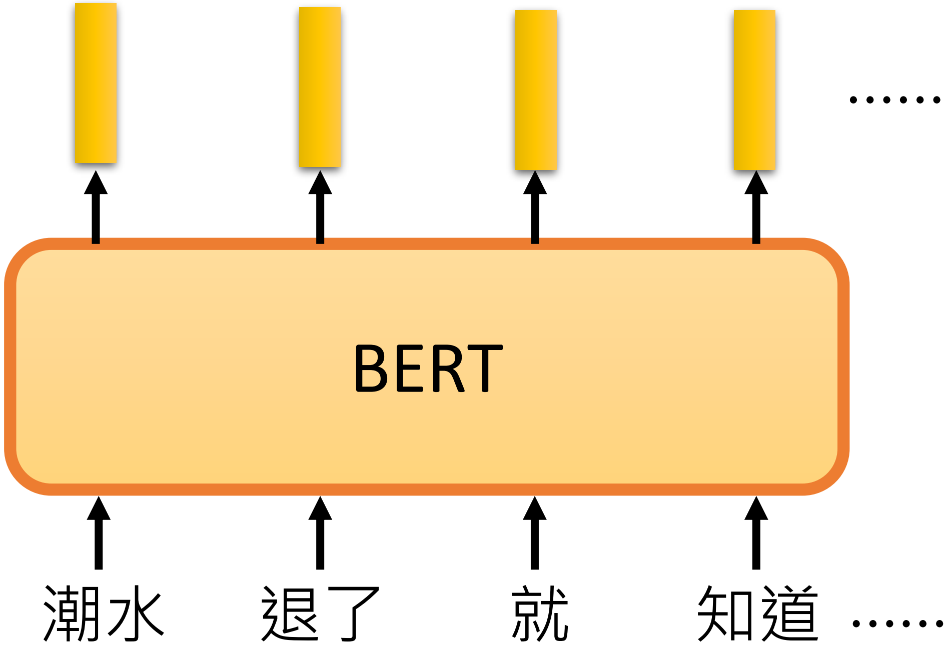
Learned with the down stream tasks

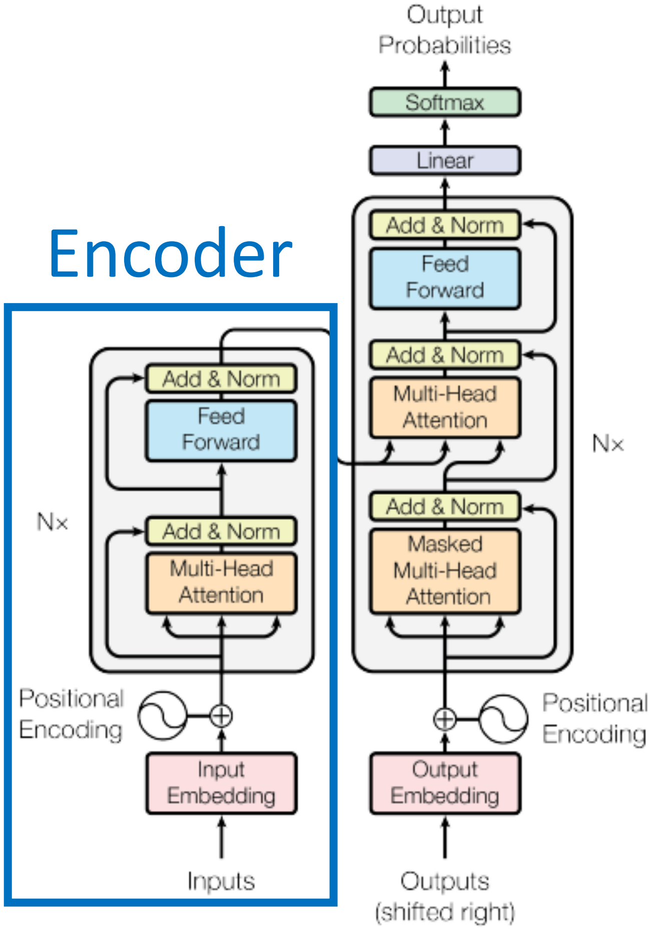


Bidirectional Encoder Representations from Transformers (BERT)

- BERT = Encoder of Transformer

Learned from a large amount of text without annotation

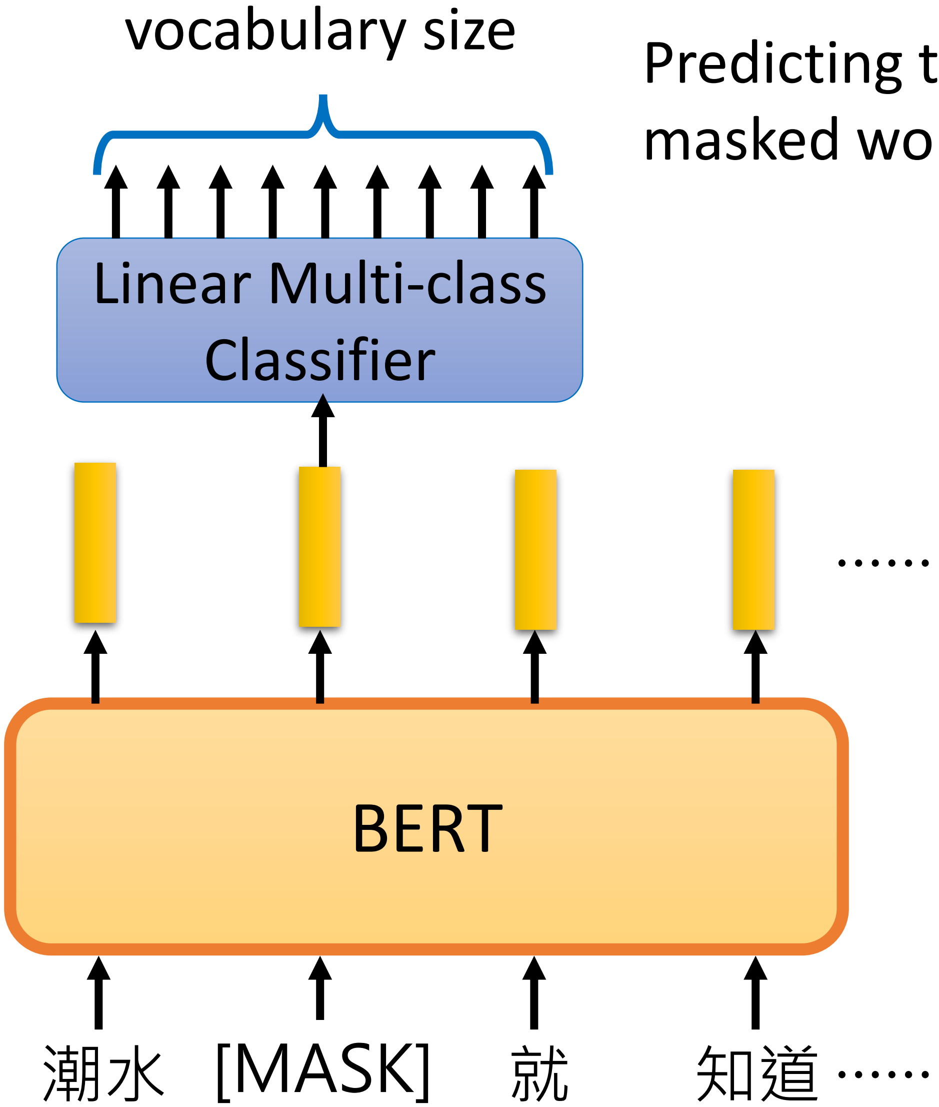



Encoder

Training of BERT

- Approach 1:

  Masked LM



vocabulary size

Predicting the masked word

Linear Multi-class Classifier

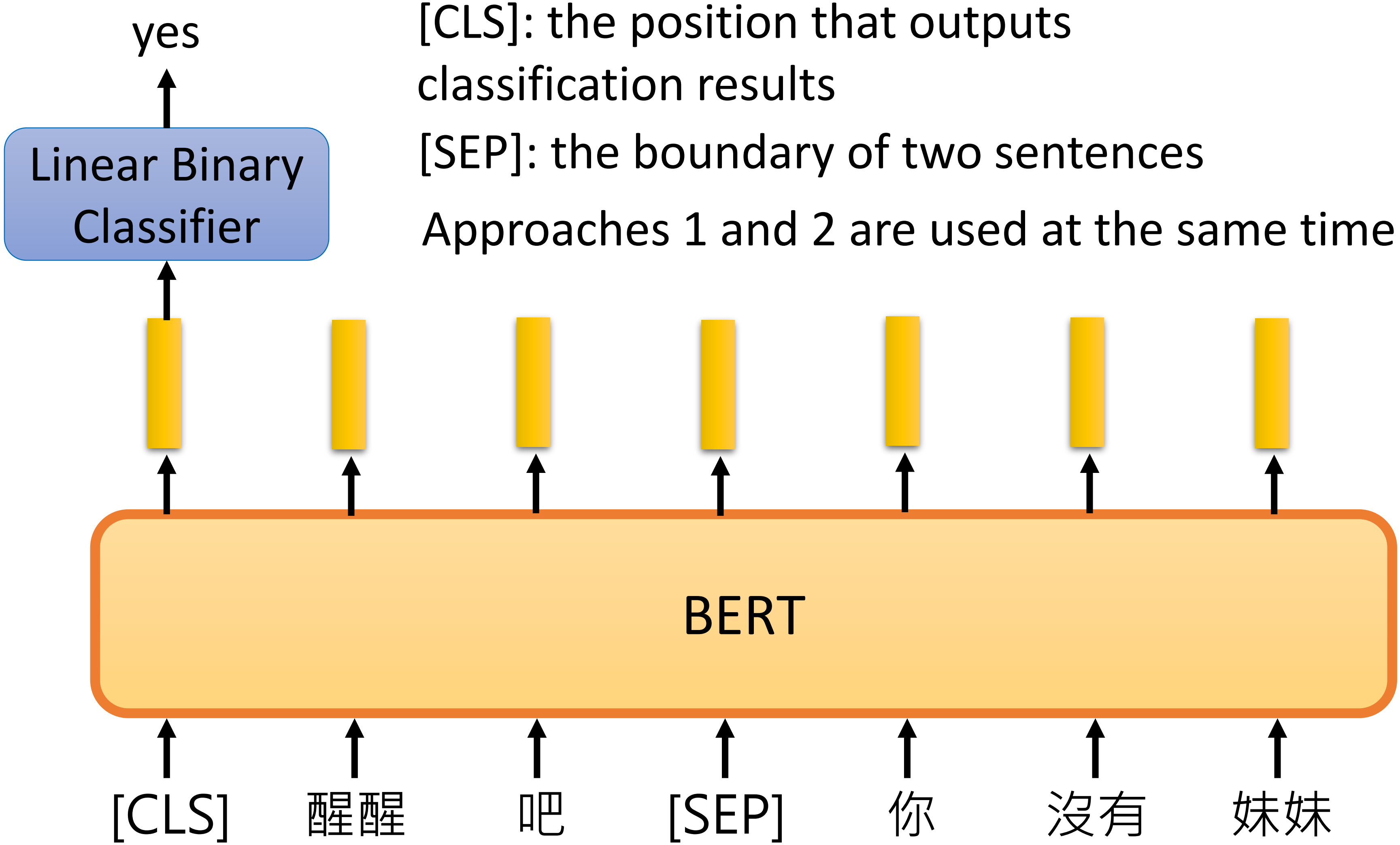
BERT

……

潮水 [MASK] 就 知道 ……

Training of BERT

Approach 2: Next Sentence Prediction



[CLS]: the position that outputs classification results

[SEP]: the boundary of two sentences

Approaches 1 and 2 are used at the same time.

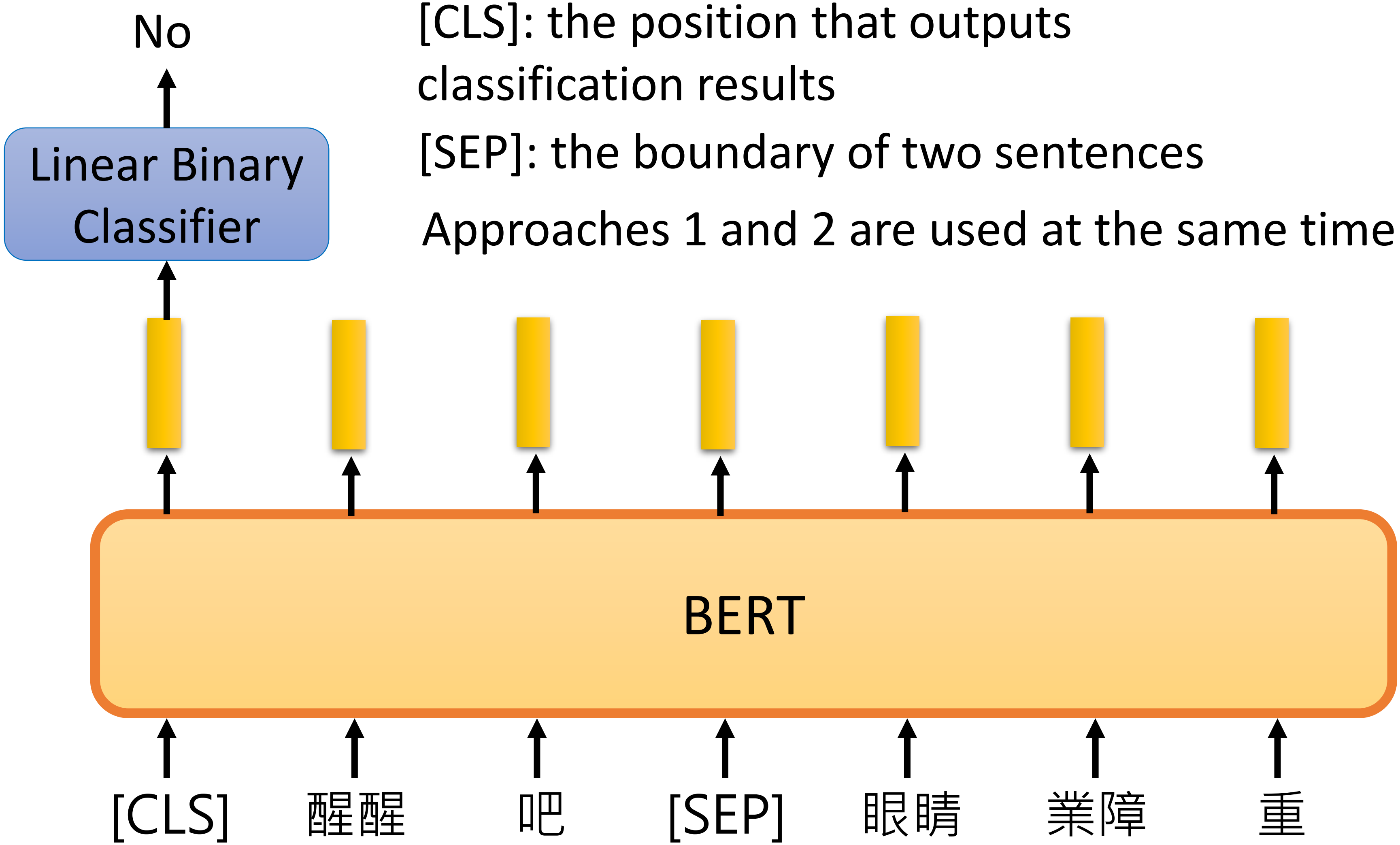
Training of BERT

Approach 2: Next Sentence Prediction

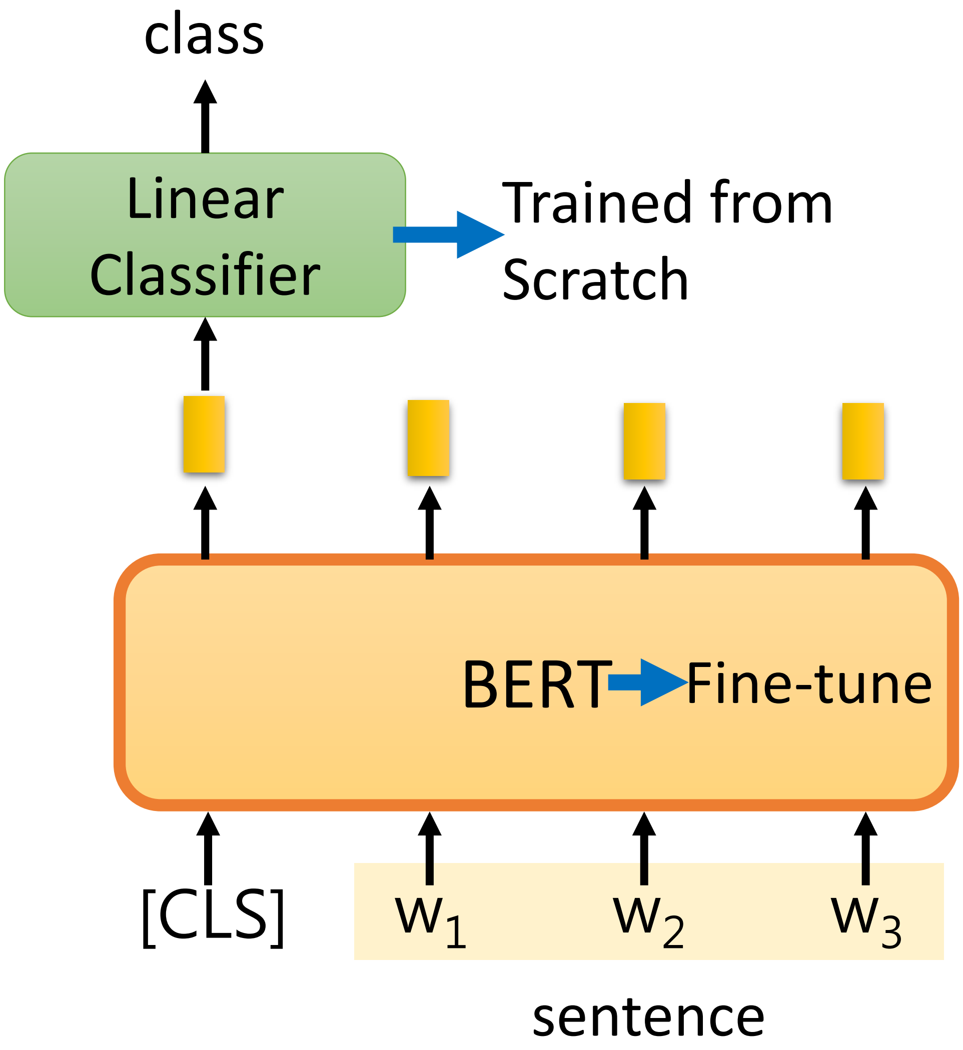
[CLS]: the position that outputs classification results

[SEP]: the boundary of two sentences

Approaches 1 and 2 are used at the same time.



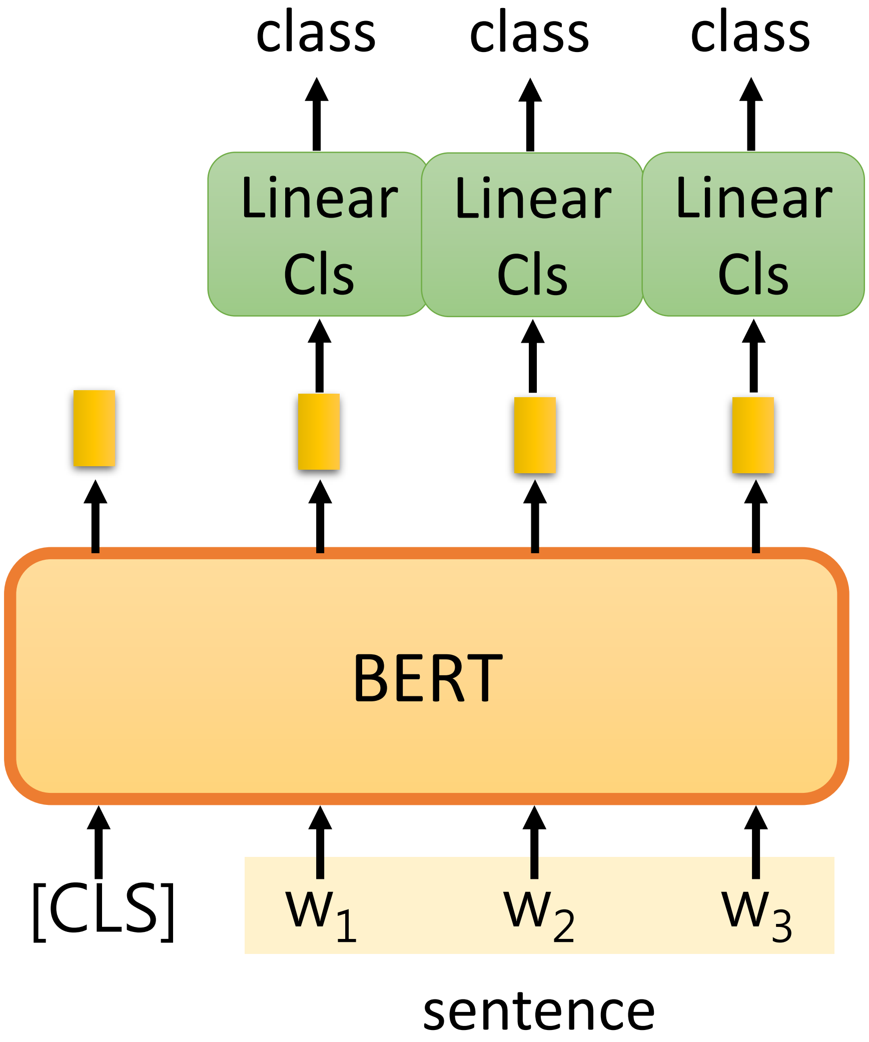
How to use BERT – Case 1



Input: single sentence, output: class

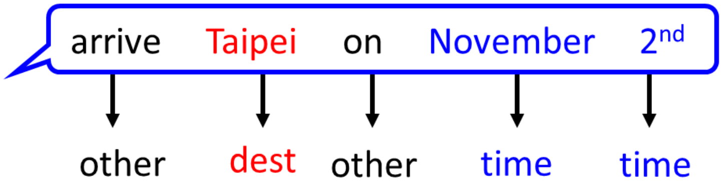
Example:
Sentiment analysis (our HW),
Document Classification

How to use BERT – Case 2



Input: single sentence,
output: class of each word

Example: Slot filling

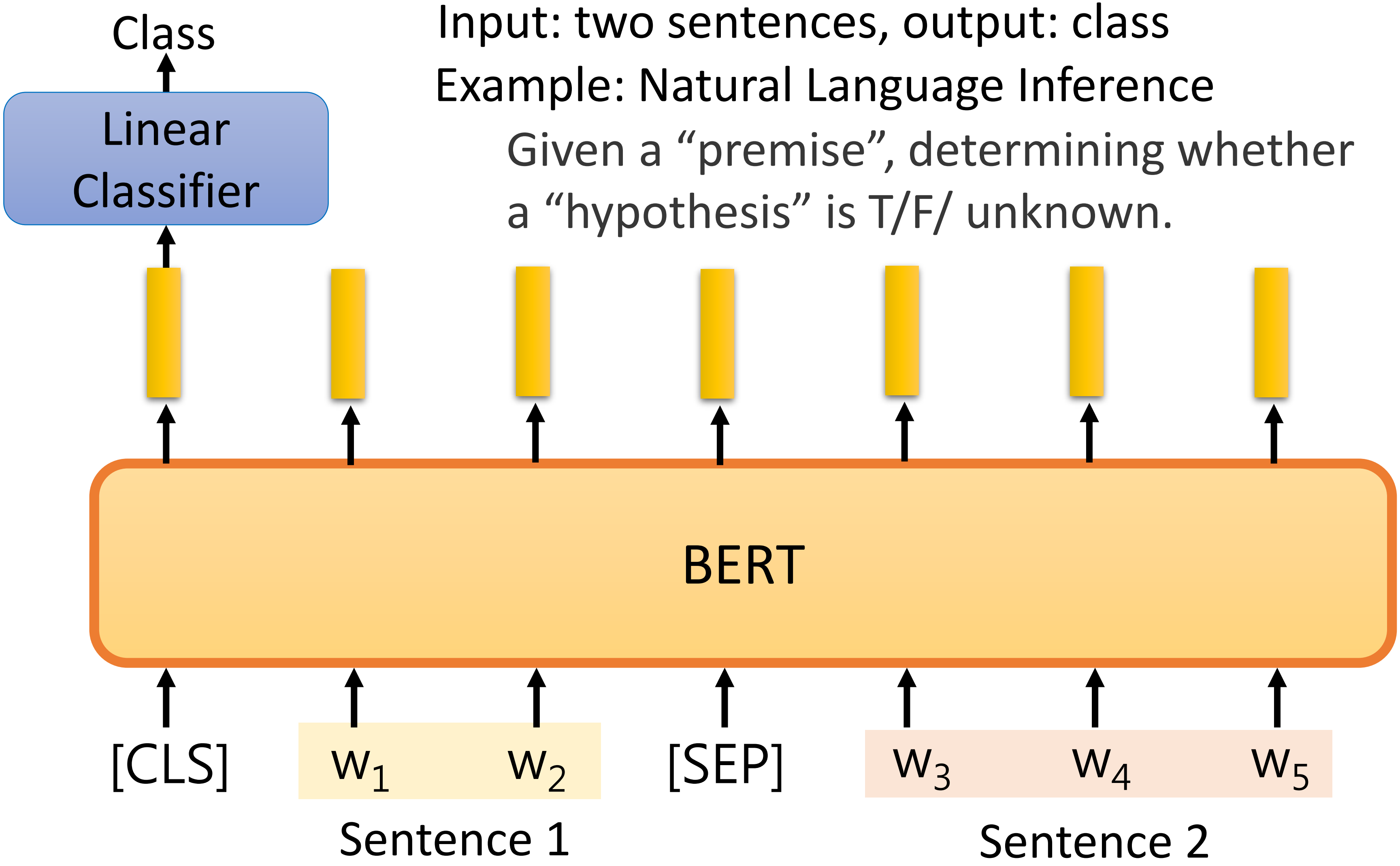


How to use BERT – Case 3

Input: two sentences, output: class

Example: Natural Language Inference

Given a “premise”, determining whether a “hypothesis” is T/F/ unknown.

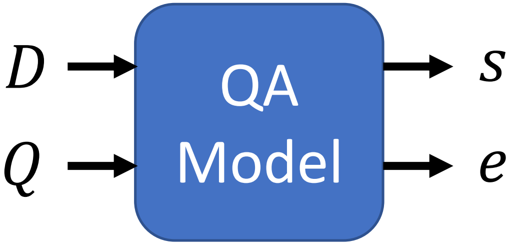


How to use BERT – Case 4

- Extraction-based Question Answering (QA) (E.g. SQuAD)

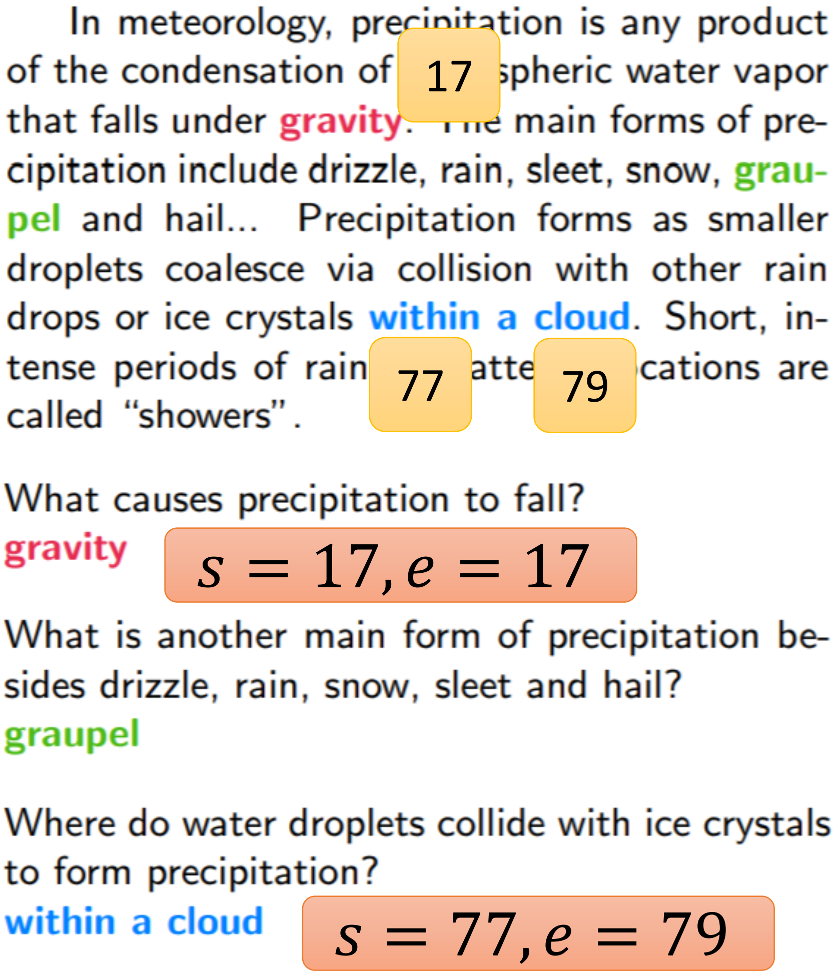
Document: $D = \{d_1, d_2, \cdots, d_N\}$

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



output: two integers (s, e)

Answer: $A = \{q_s, \cdots, q_e\}$



In meteorology, precipitation is any product of the condensation of 17 spheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain 77 atte 79 cations 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

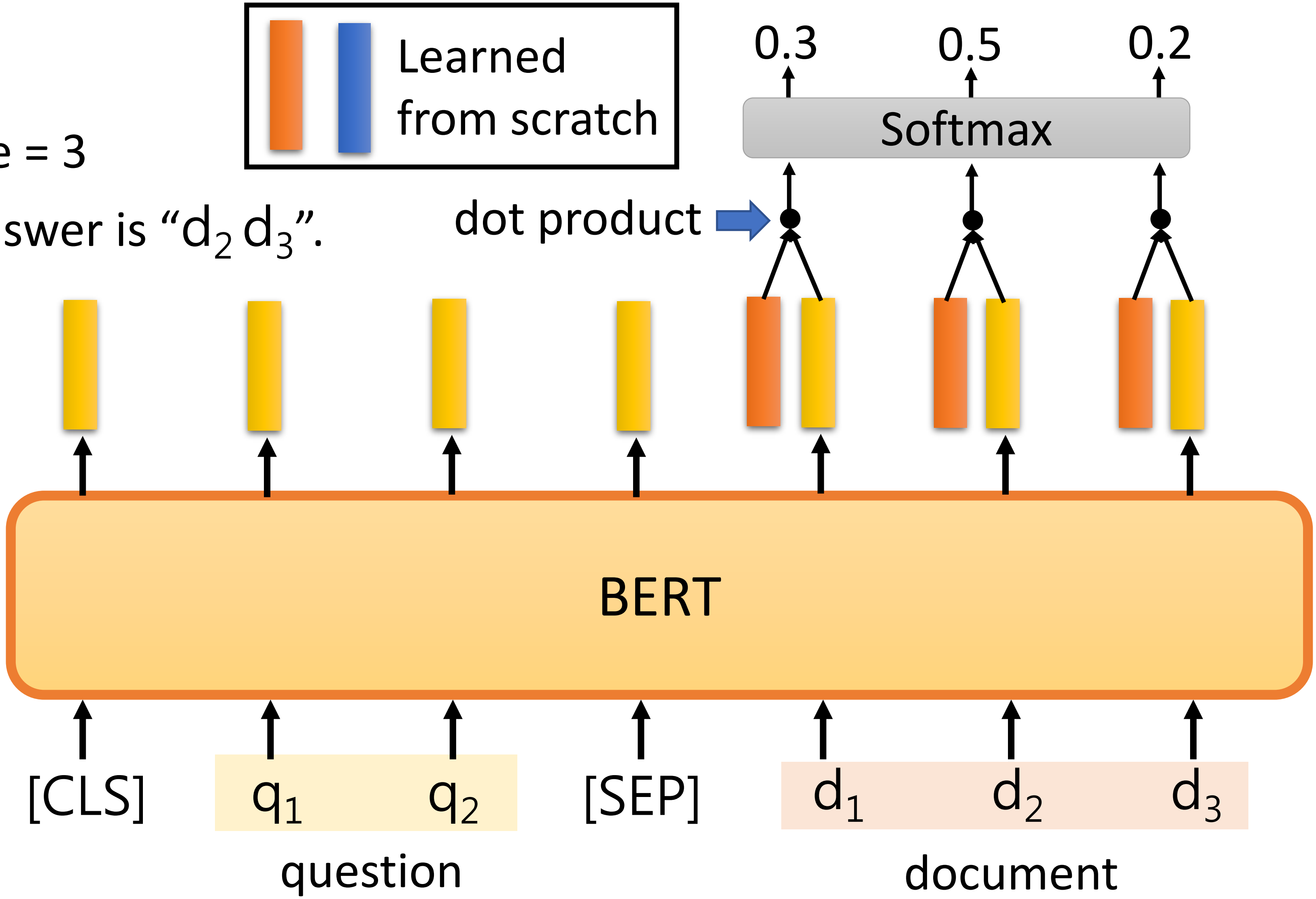
Where do water droplets collide with ice crystals to form precipitation?

within a cloud $s = 77, e = 79$

How to use BERT – Case 4

s = 2, e = 3

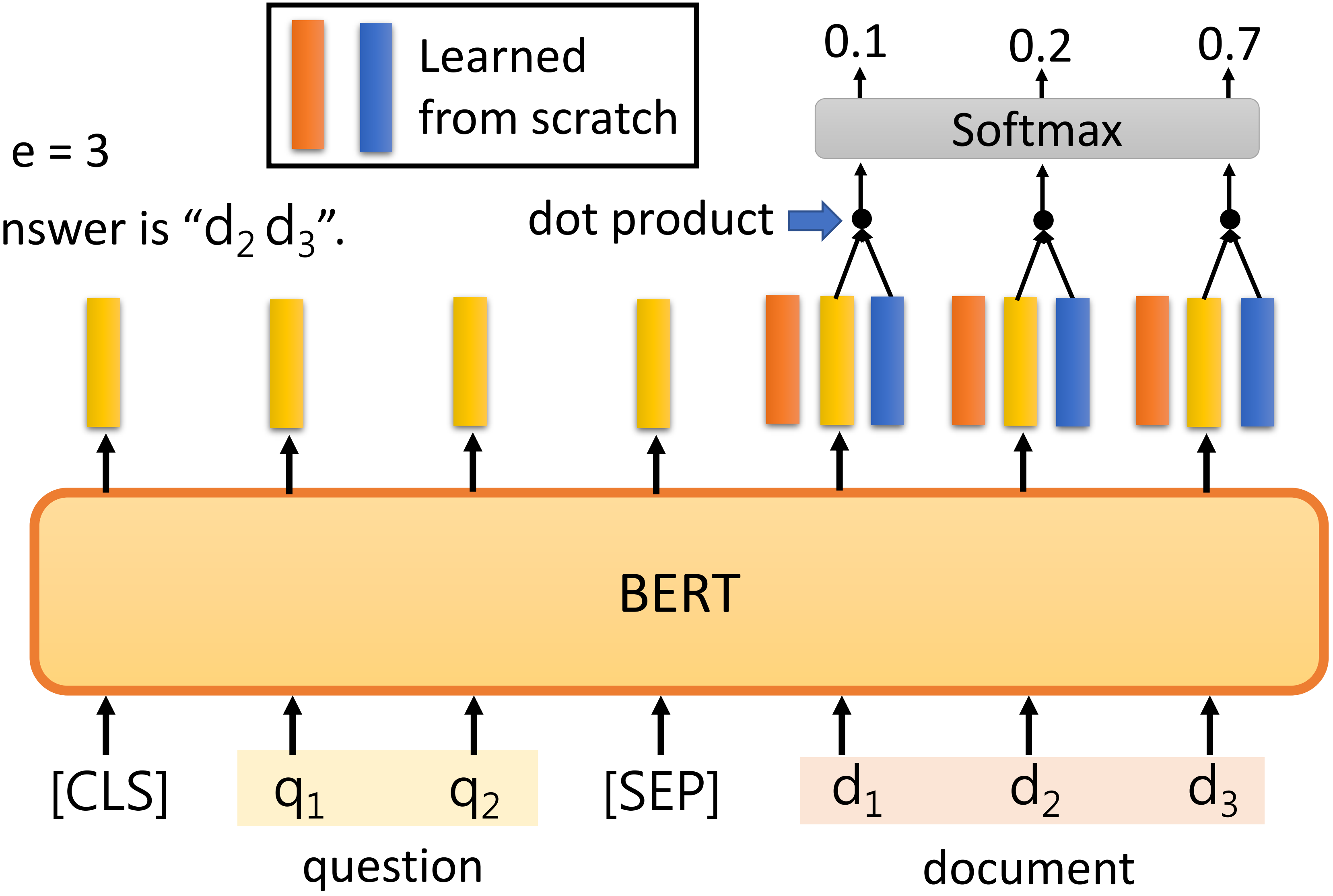
The answer is “d_2 d_3”.



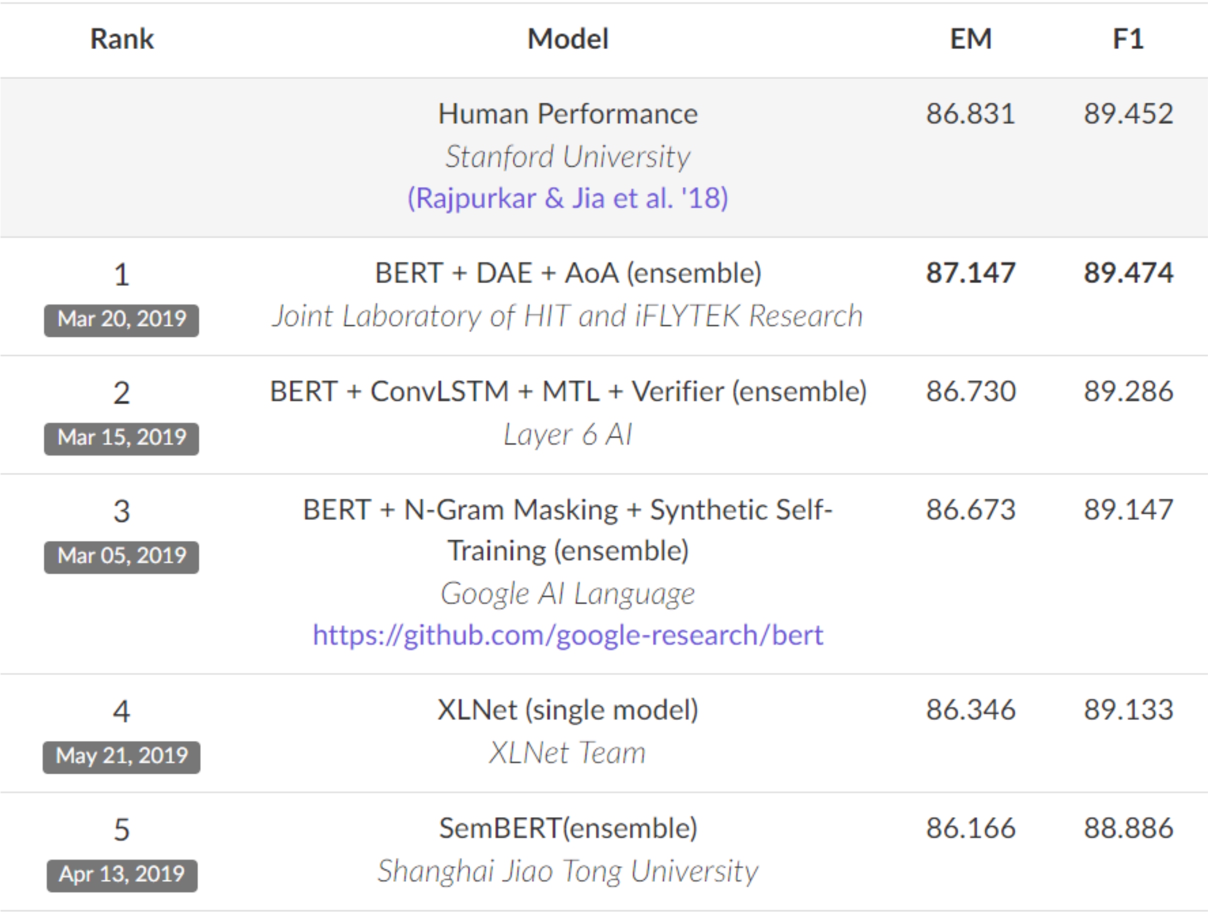
How to use BERT – Case 4

s = 2, e = 3

The answer is “d_2 d_3”.



BERT 屠榜 ……

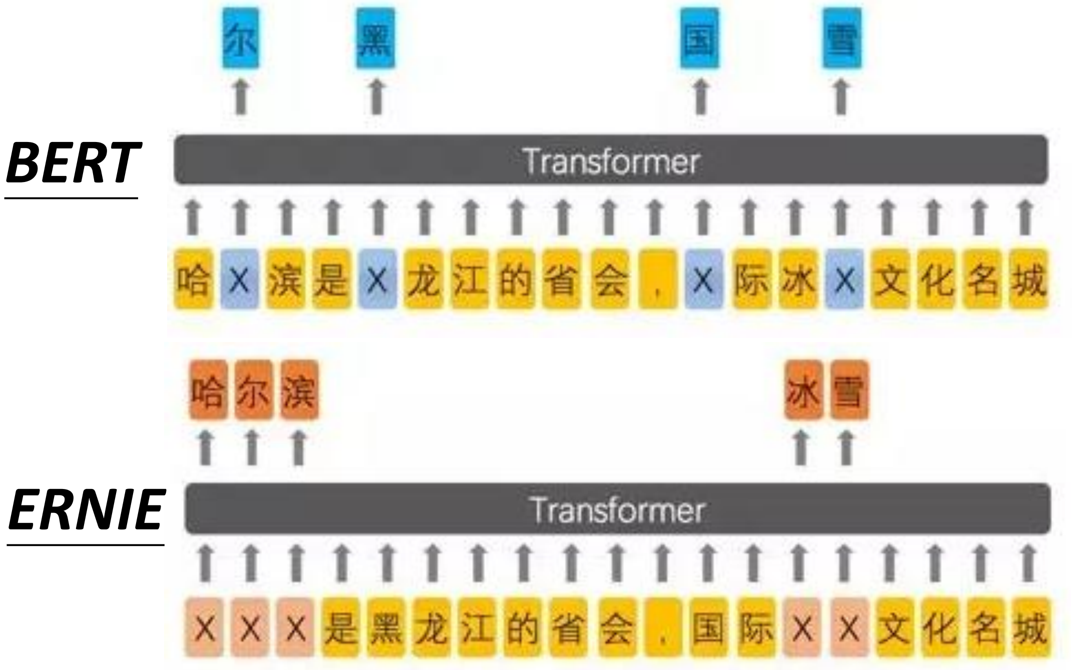


Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	**87.147**	**89.474**
2 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
3 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	86.673	89.147
4 May 21, 2019	XLNet (single model) *XLNet Team*	86.346	89.133
5 Apr 13, 2019	SemBERT(ensemble) *Shanghai Jiao Tong University*	86.166	88.886

SQuAD 2.0

Enhanced Representation through Knowledge Integration (ERNIE)

- Designed for Chinese


尔
黑
国
雪
BERT
Transformer
哈 X 滨 是 X 龙 江 的 省 会 ， X 际 冰 X 文 化 名 城
哈 尔 滨
冰 雪
ERNIE
Transformer
X X X 是 黑 龙 江 的 省 会 ， 国 际 X X 文 化 名 城

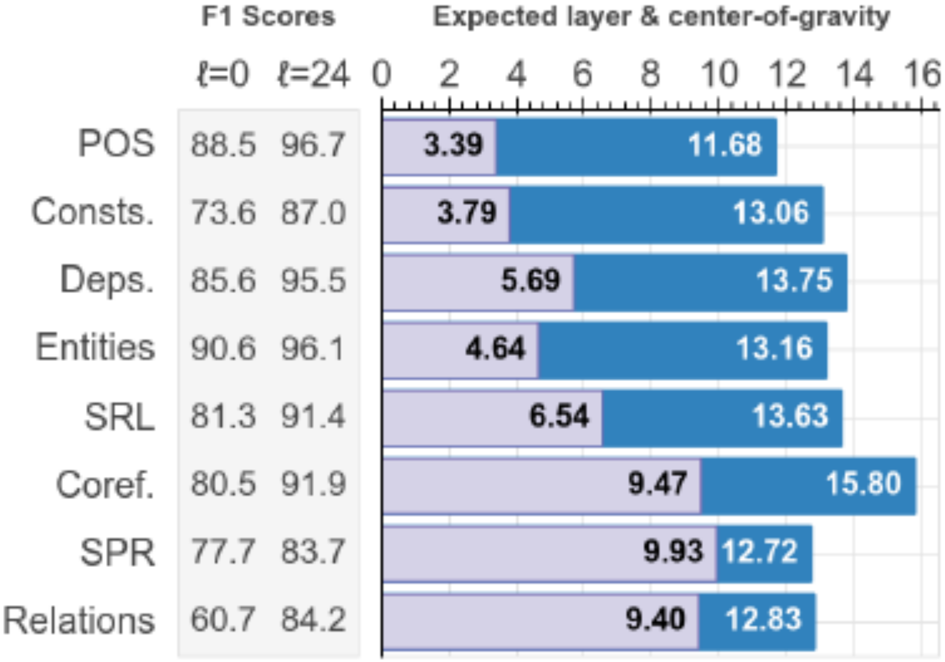
Source of image:
https://zhuanlan.zhihu.com/p/59436589

https://arxiv.org/abs/1904.09223

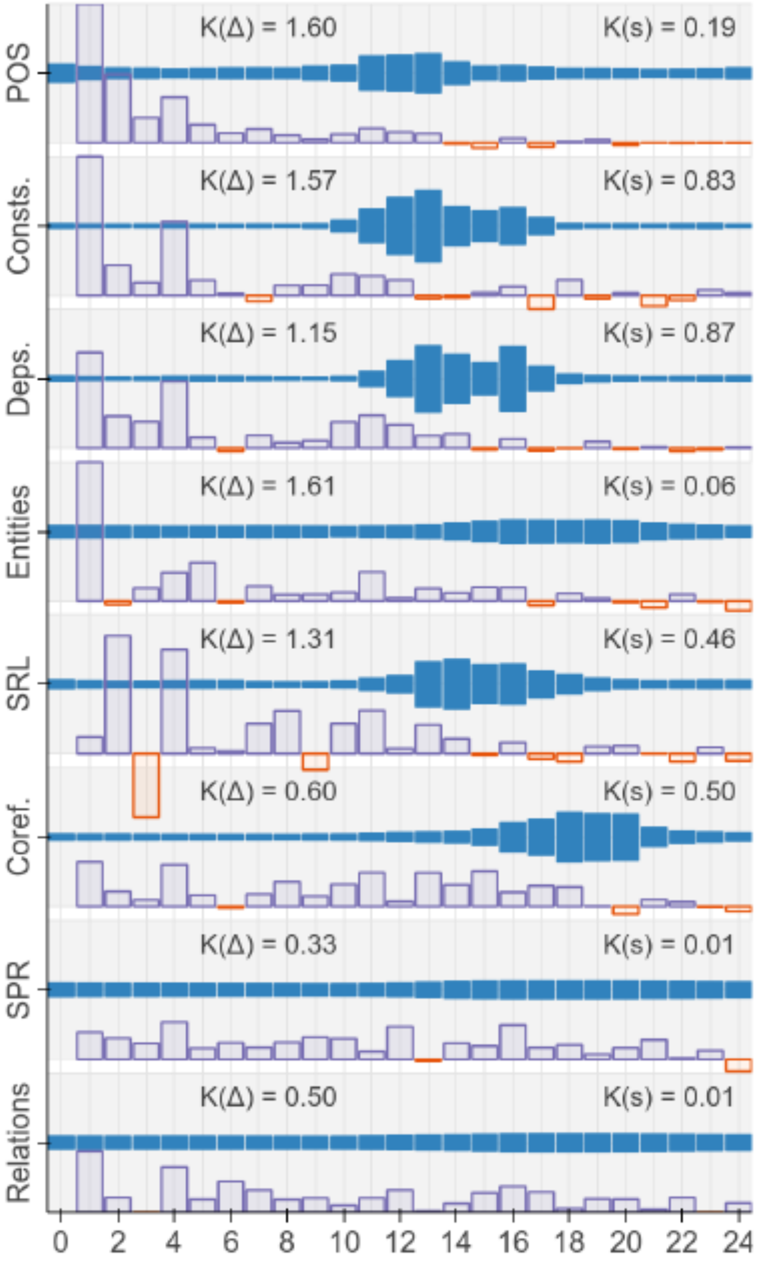
What does BERT learn?

https://arxiv.org/abs/1905.05950

https://openreview.net/pdf?id=SJzSgnRcKX



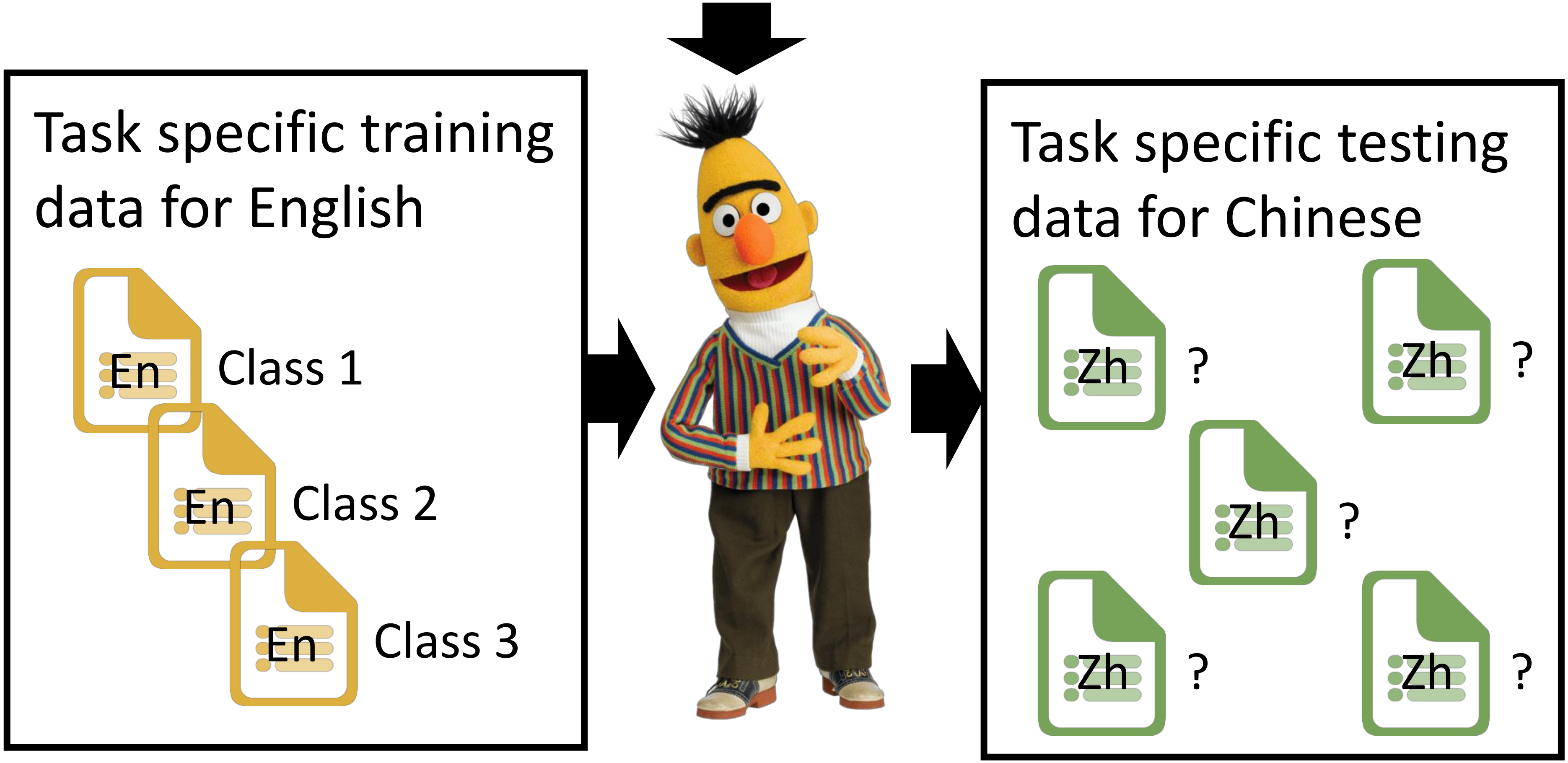
	F1 Scores		Expected layer & center-of-gravity	
	ℓ=0	ℓ=24		
POS	88.5	96.7	3.39	11.68
Consts.	73.6	87.0	3.79	13.06
Deps.	85.6	95.5	5.69	13.75
Entities	90.6	96.1	4.64	13.16
SRL	81.3	91.4	6.54	13.63
Coref.	80.5	91.9	9.47	15.80
SPR	77.7	83.7	9.93	12.72
Relations	60.7	84.2	9.40	12.83



https://arxiv.org/abs/1904.09077

Multilingual BERT

Trained on 104 languages

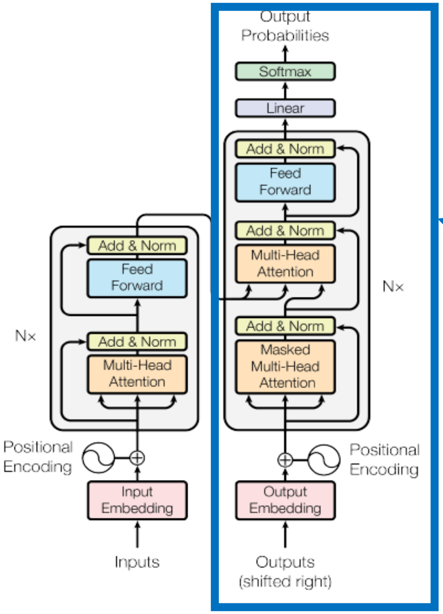


Task specific training data for English

Task specific testing data for Chinese

https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Generative Pre-Training (GPT)



Transformer Decoder

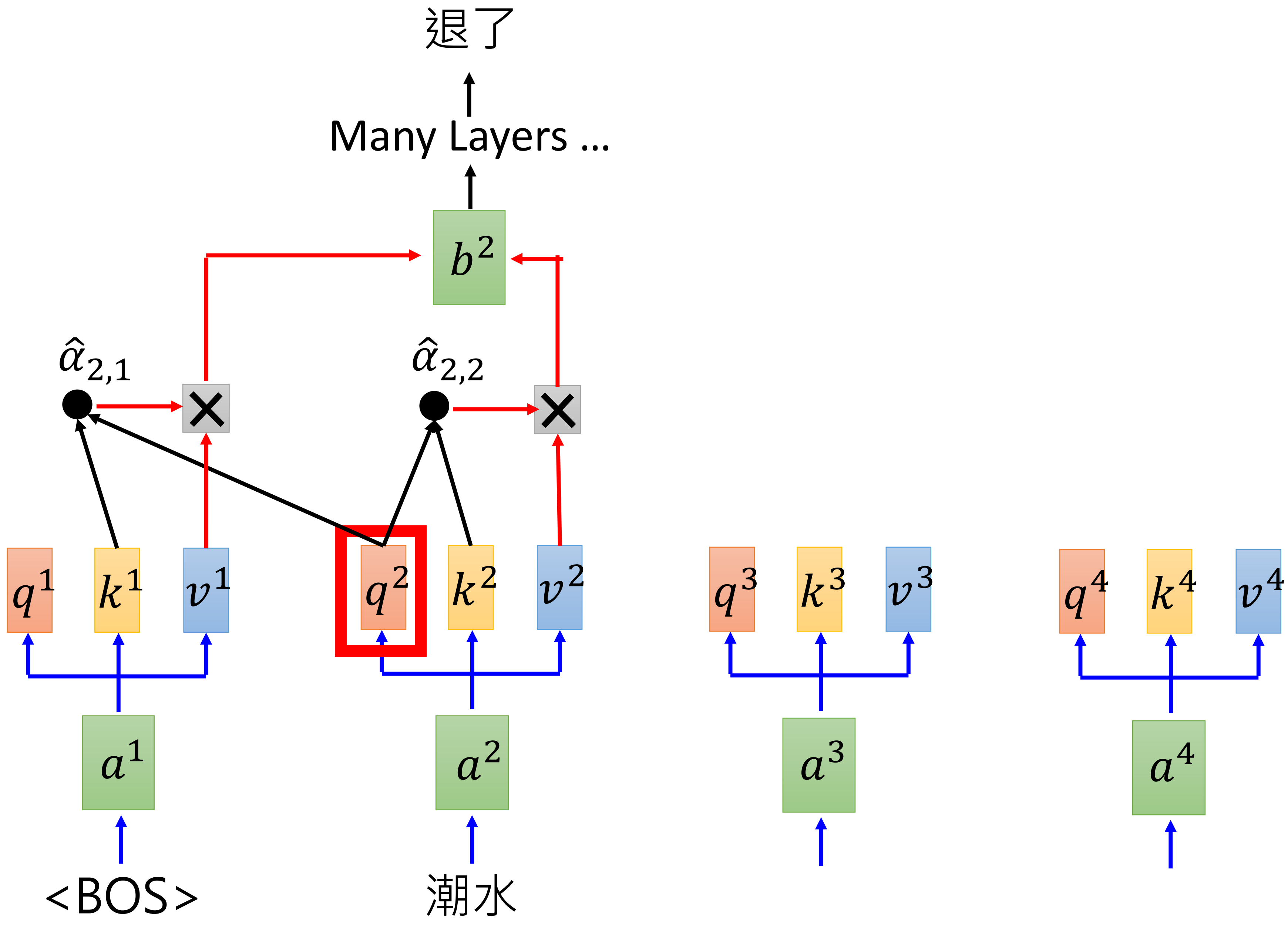
BERT (340M)

ELMO (94M)

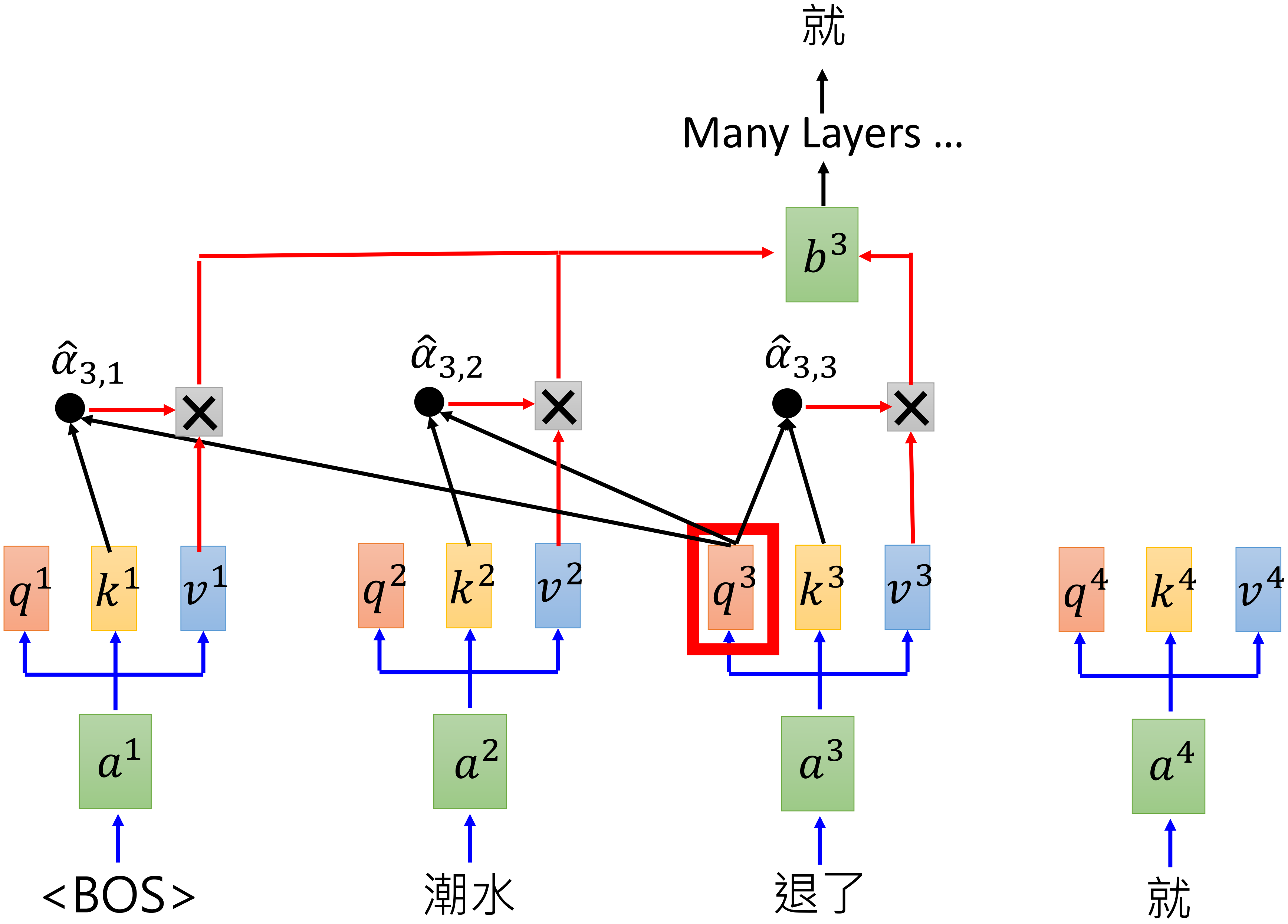
GPT-2 (1542M)

Source of image: https://huaban.com/pins/1714071707/

Generative Pre-Training (GPT)



Generative Pre-Training (GPT)

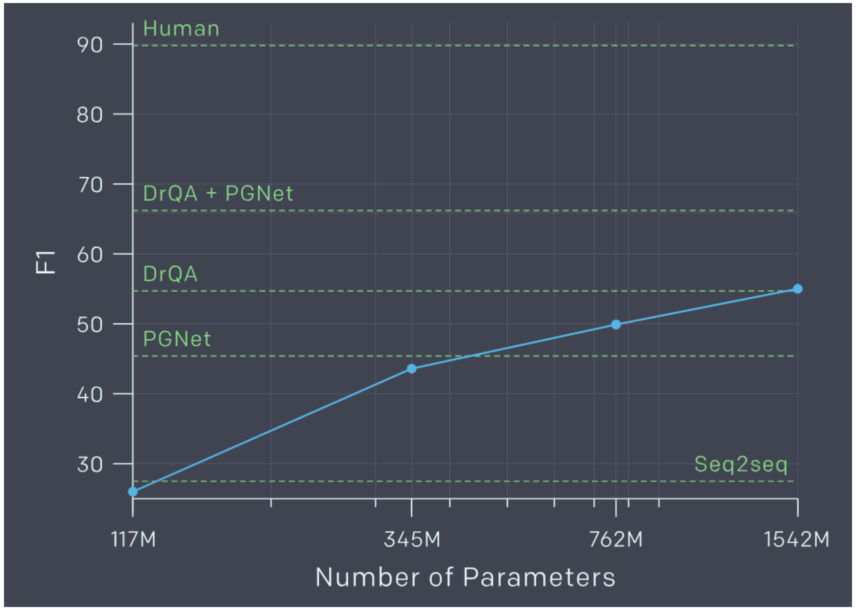


Zero-shot Learning?

- ***Reading Comprehension***

$d_1, d_2, \cdots, d_N,$
”Q:”, $q_1, q_2, \cdots, q_N,$
“A:”

CoQA



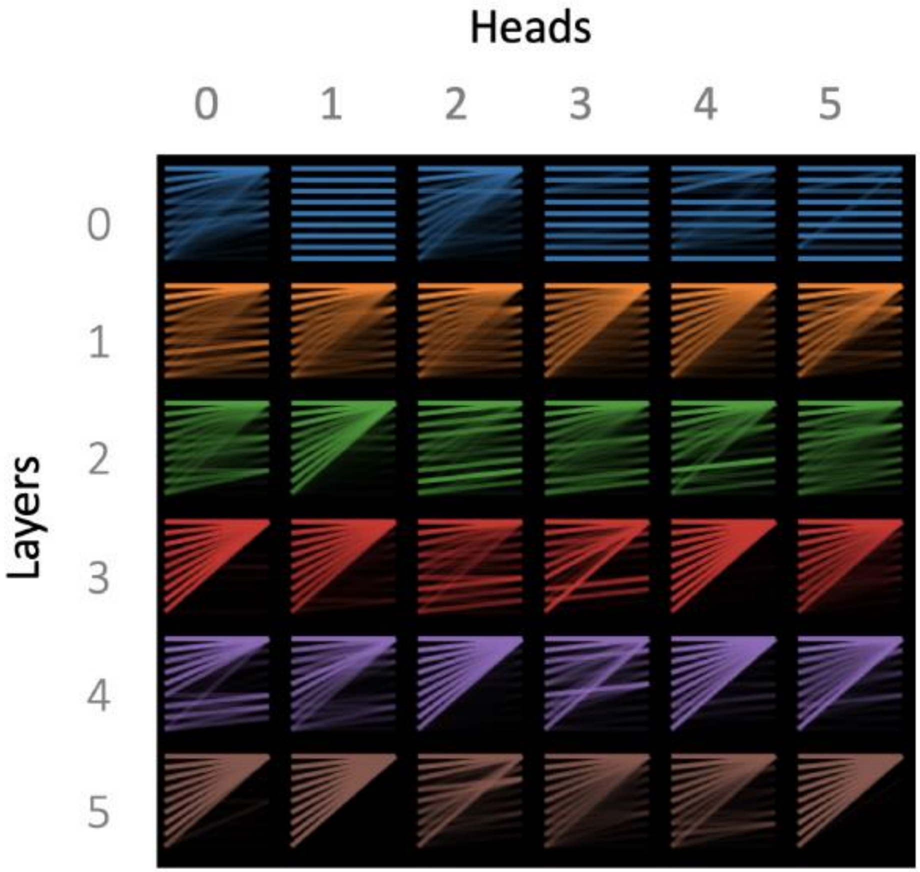
- ***Summarization*** $d_1, d_2, \cdots, d_N,$”TL;DR:”

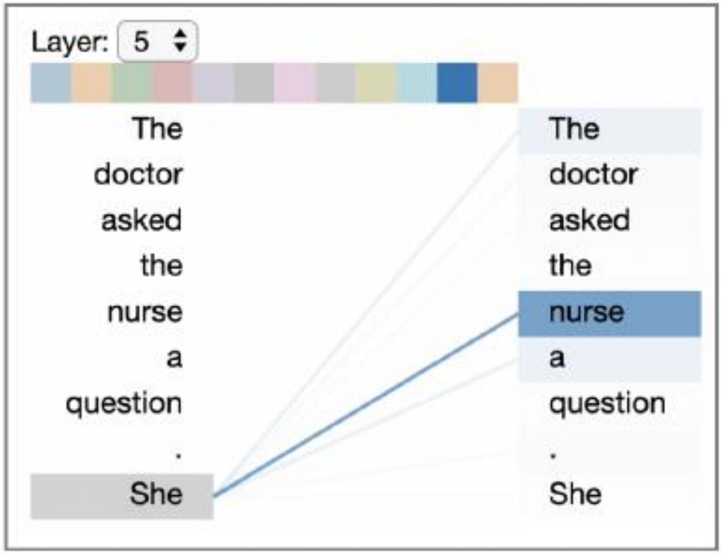
- ***Translation***

English sentence 1 = French sentence 1
English sentence 2 = French sentence 2
English sentence 3 =

Visualization

https://arxiv.org/abs/1904.02679
(The results below are from GPT-2)

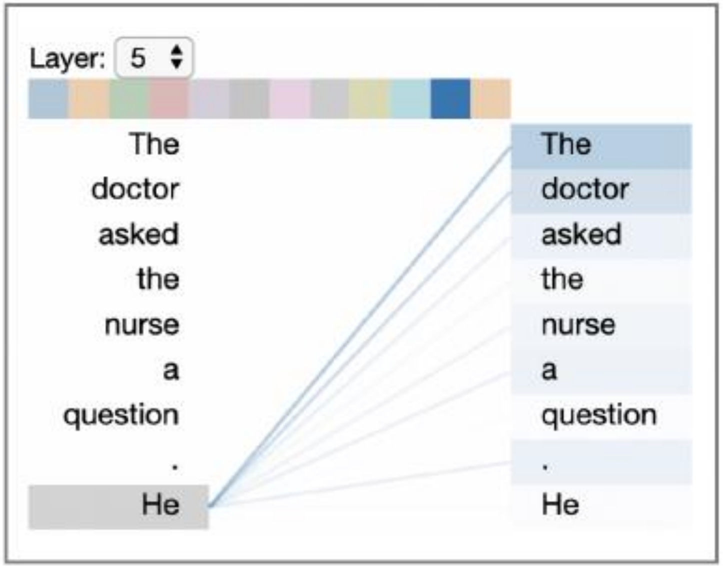




Layer: 5

The doctor asked the nurse a question . She

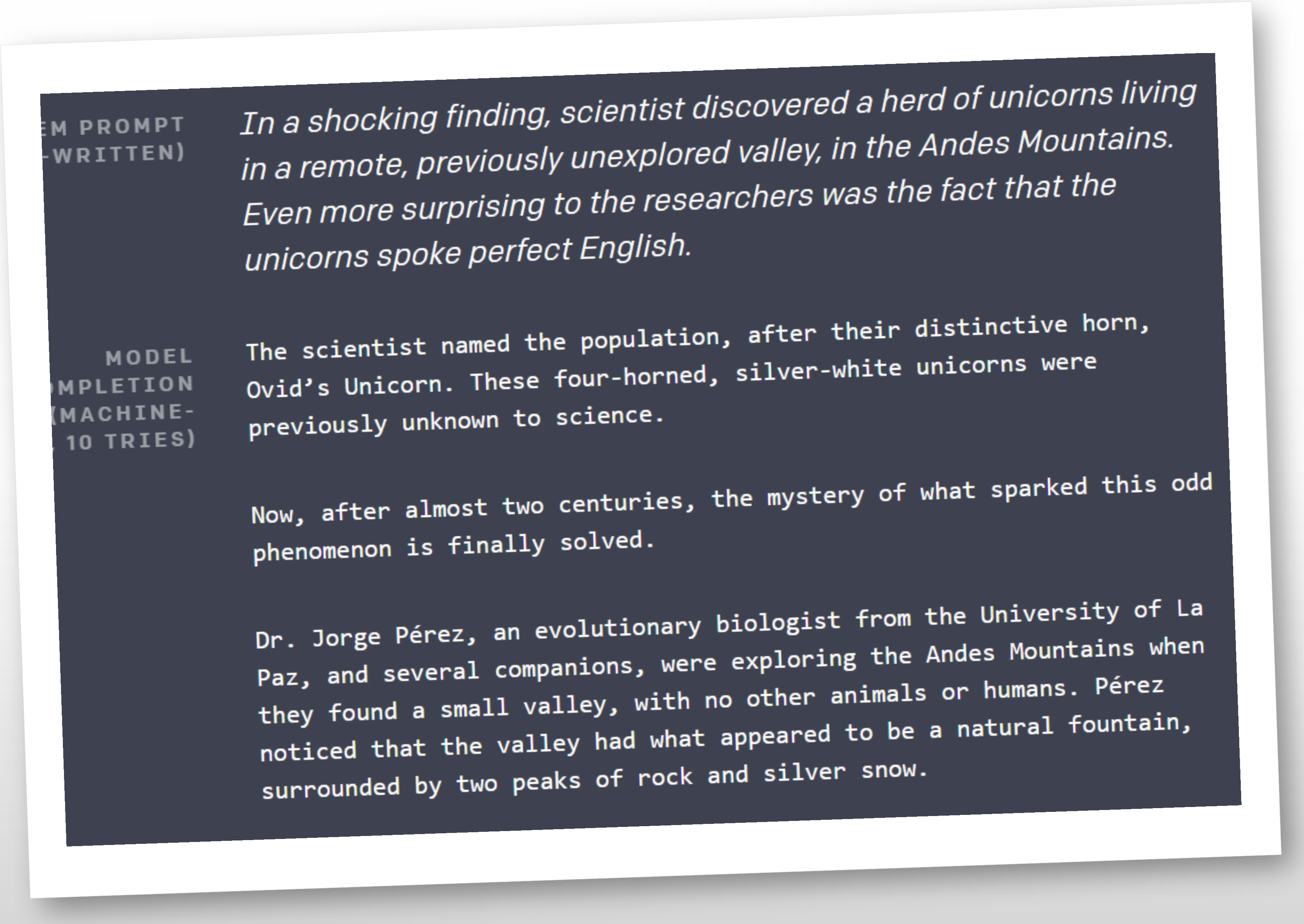
The doctor asked the nurse a question . She



Layer: 5

The doctor asked the nurse a question . He

The doctor asked the nurse a question . He



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.

https://talktotransformer.com/

Can BERT speak?

- Unified Language Model Pre-training for Natural Language Understanding and Generation
    - https://arxiv.org/abs/1905.03197
- BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model
    - https://arxiv.org/abs/1902.04094
- Insertion Transformer: Flexible Sequence Generation via Insertion Operations
    - https://arxiv.org/abs/1902.03249
- Insertion-based Decoding with automatically Inferred Generation Order
    - https://arxiv.org/abs/1902.01370