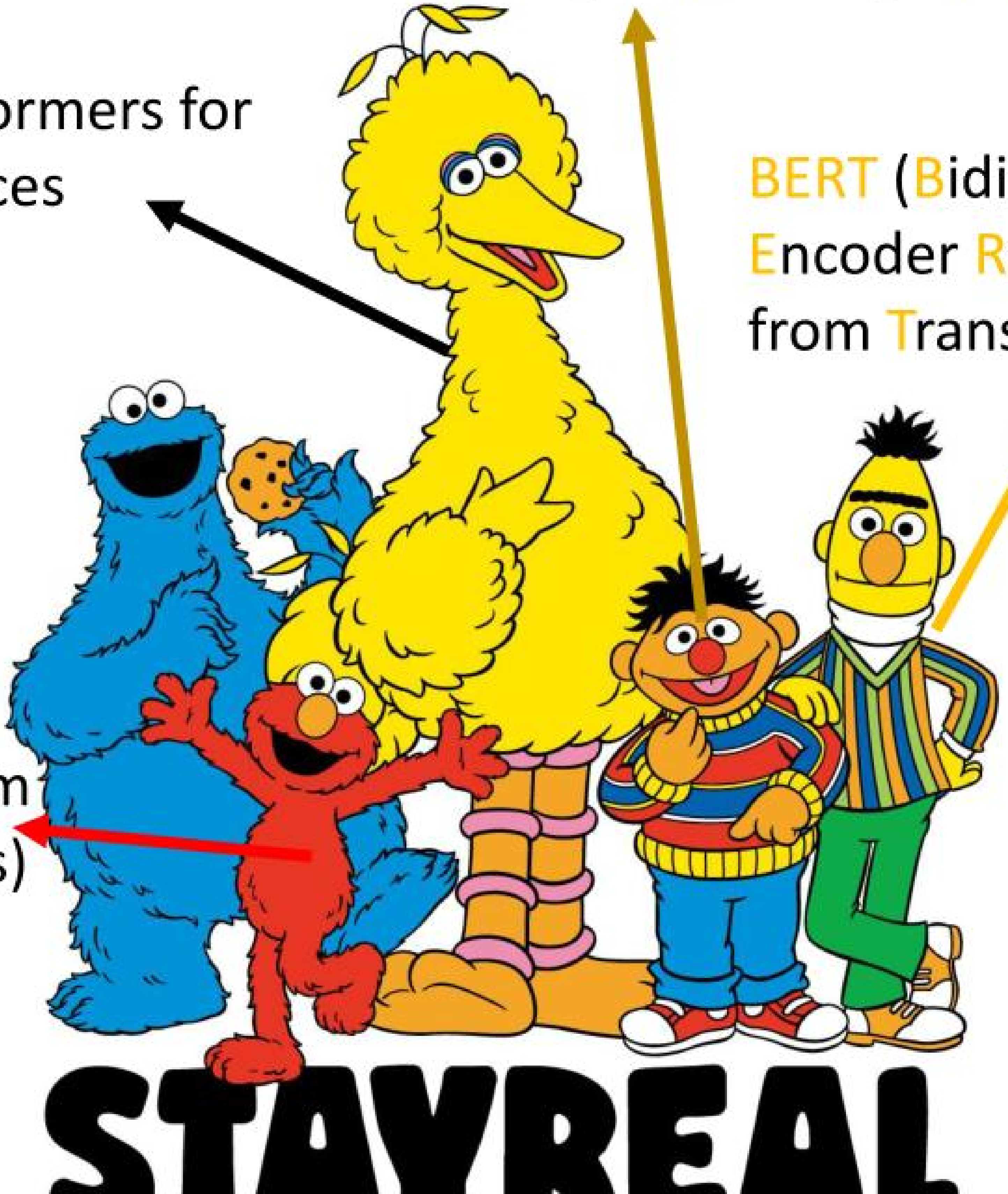
人工智能技术及应用

Artificial Intelligence and Application

Self-Supervised Learning



ERNIE (Enhanced Representation through Knowledge Integration)

Big Bird: Transformers for Longer Sequences

BERT (Bidirectional Encoder Representations from Transformers)

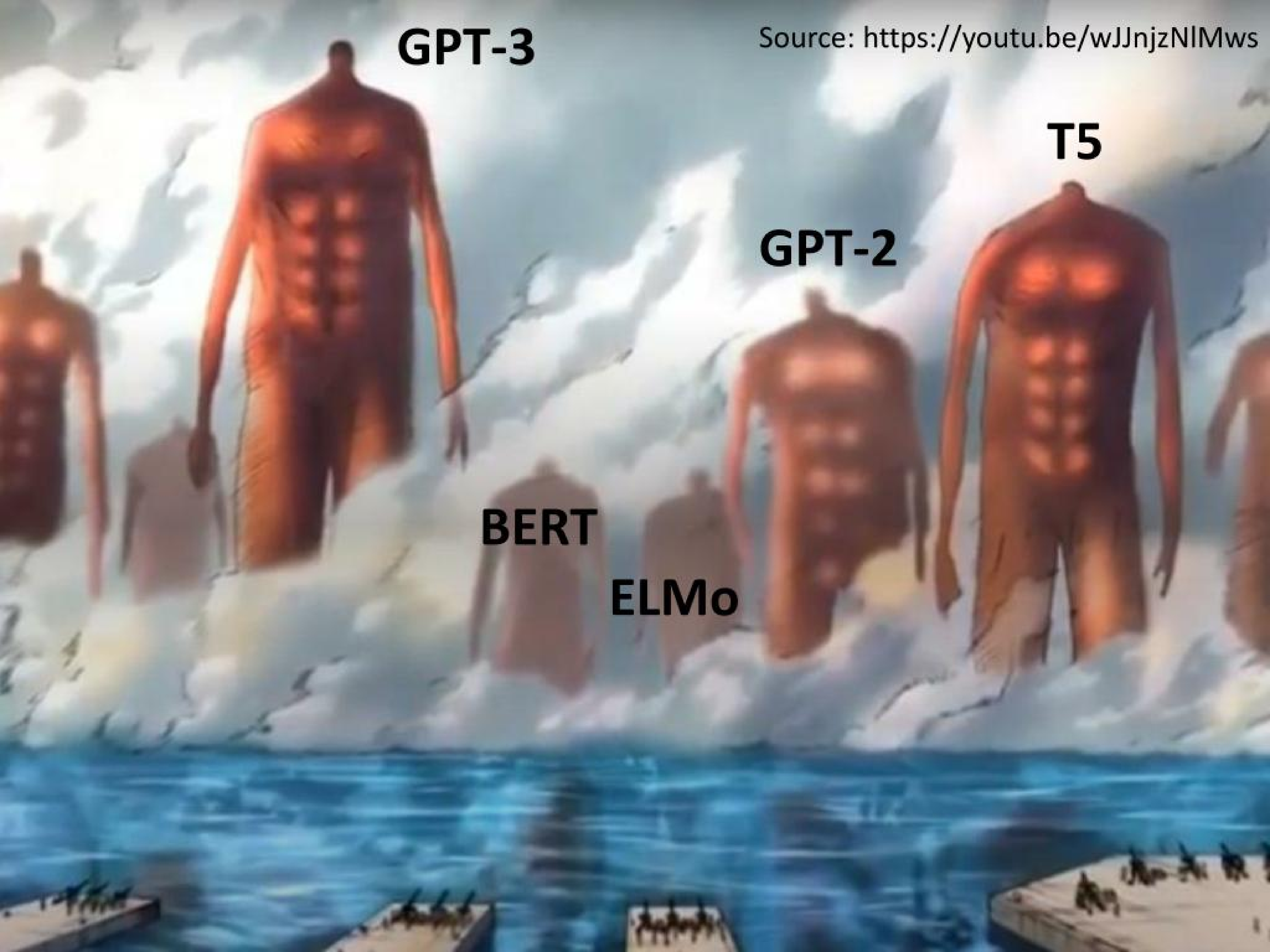
ELMo (Embeddings from Language Models)

BERT

340M parameters

Bertolt Hoover

Source of image: https://leemeng.tw/attack_on_bert_transfer_learning_in_nlp.html



Source: https://youtu.be/wJJnjzNlMws

The models become larger and larger …

ELMO
(94M)

BERT
(340M)

GPT-2
(1542M)

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

The models become larger and larger …

GPT-3 is **10** times larger than Turing NLG. (175B)

GPT-2

Megatron (8B)

T5 (11B)

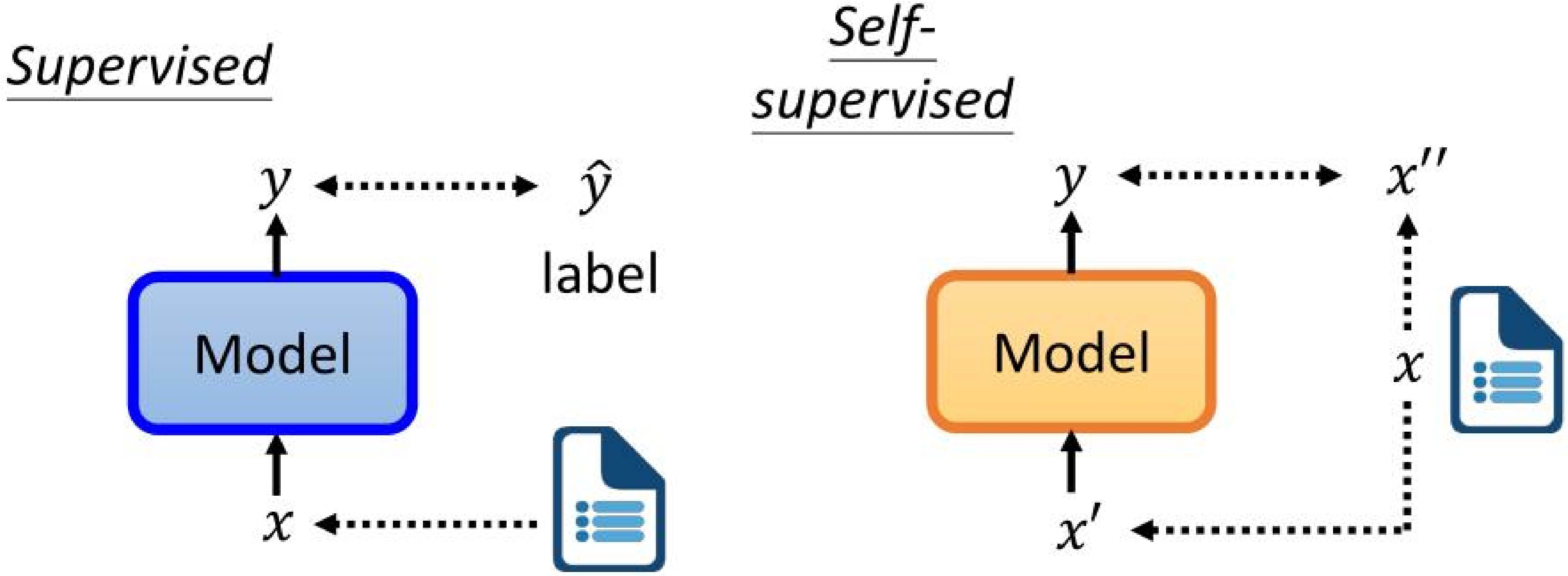
Turing NLG (17B)

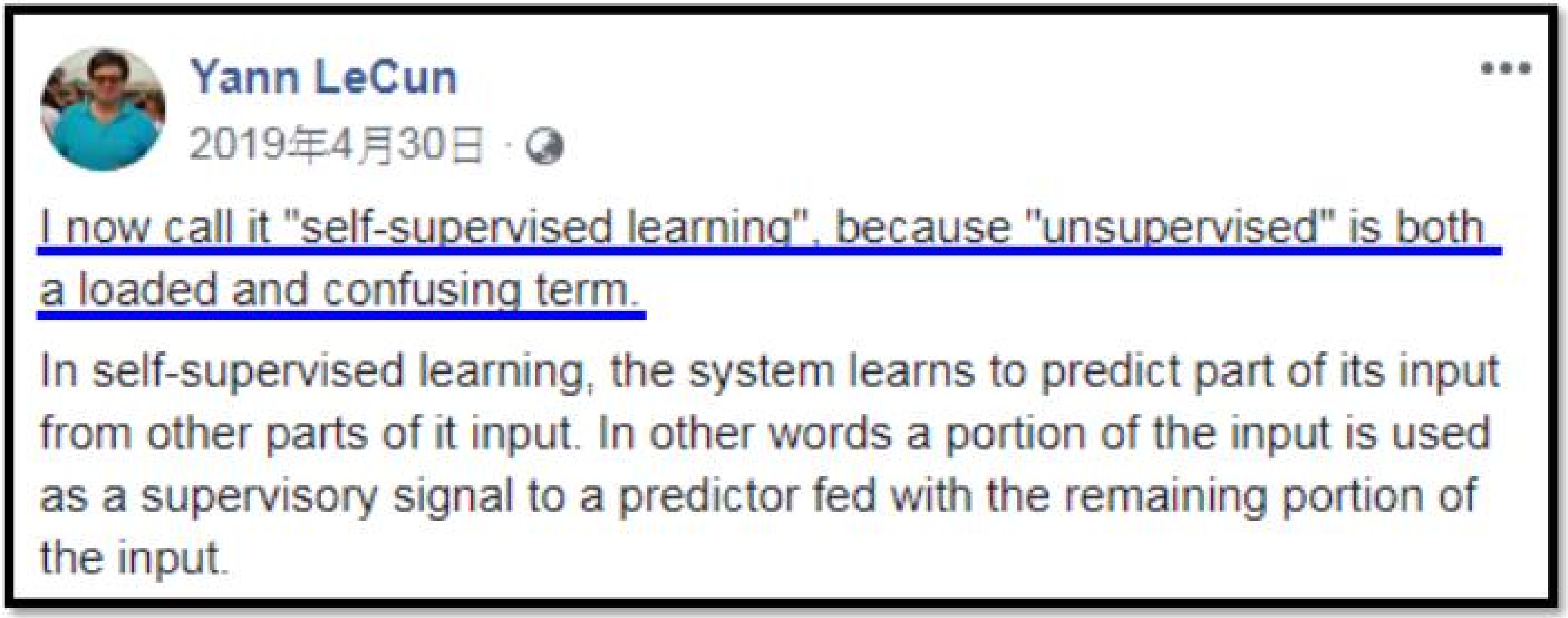
Outline

BERT series

GPT series

Self-supervised Learning





Yann LeCun

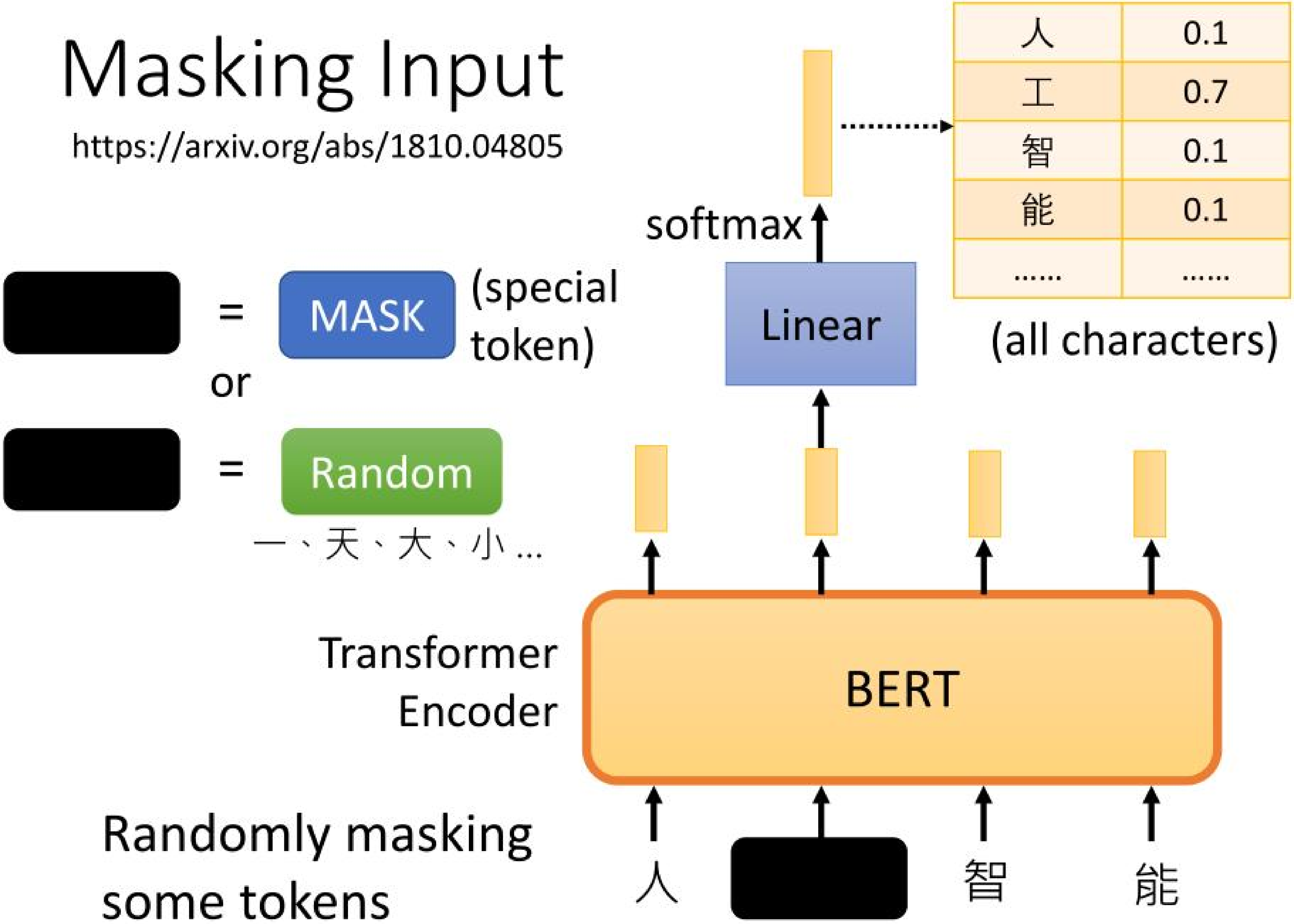
2019年4月30日 ·

I now call it "self-supervised learning", because "unsupervised" is both a loaded and confusing term.

In self-supervised learning, the system learns to predict part of its input from other parts of it input. In other words a portion of the input is used as a supervisory signal to a predictor fed with the remaining portion of the input.

Masking Input

https://arxiv.org/abs/1810.04805

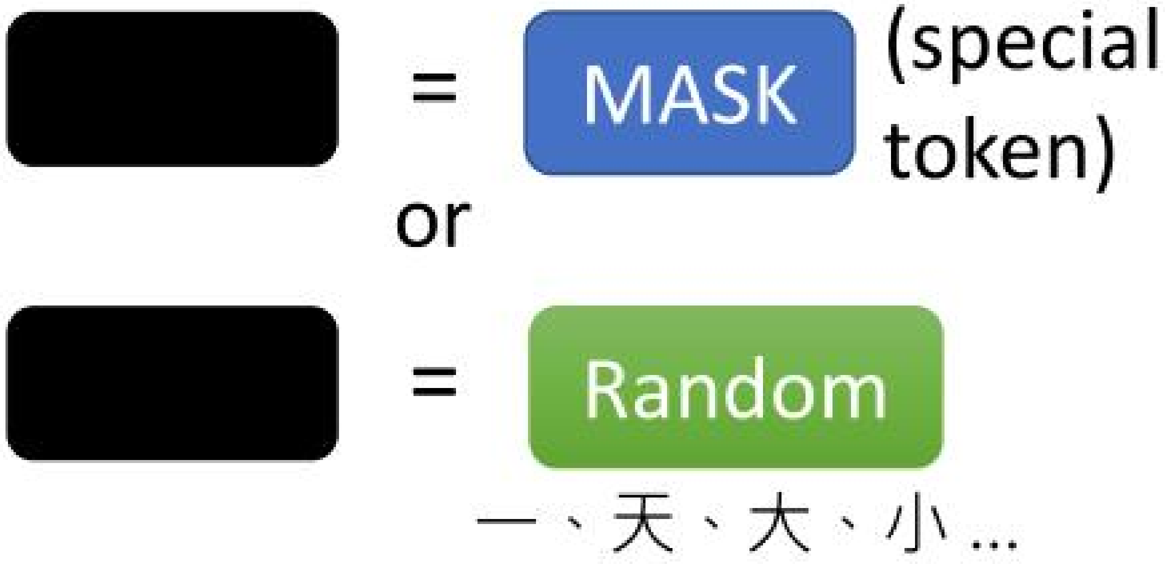


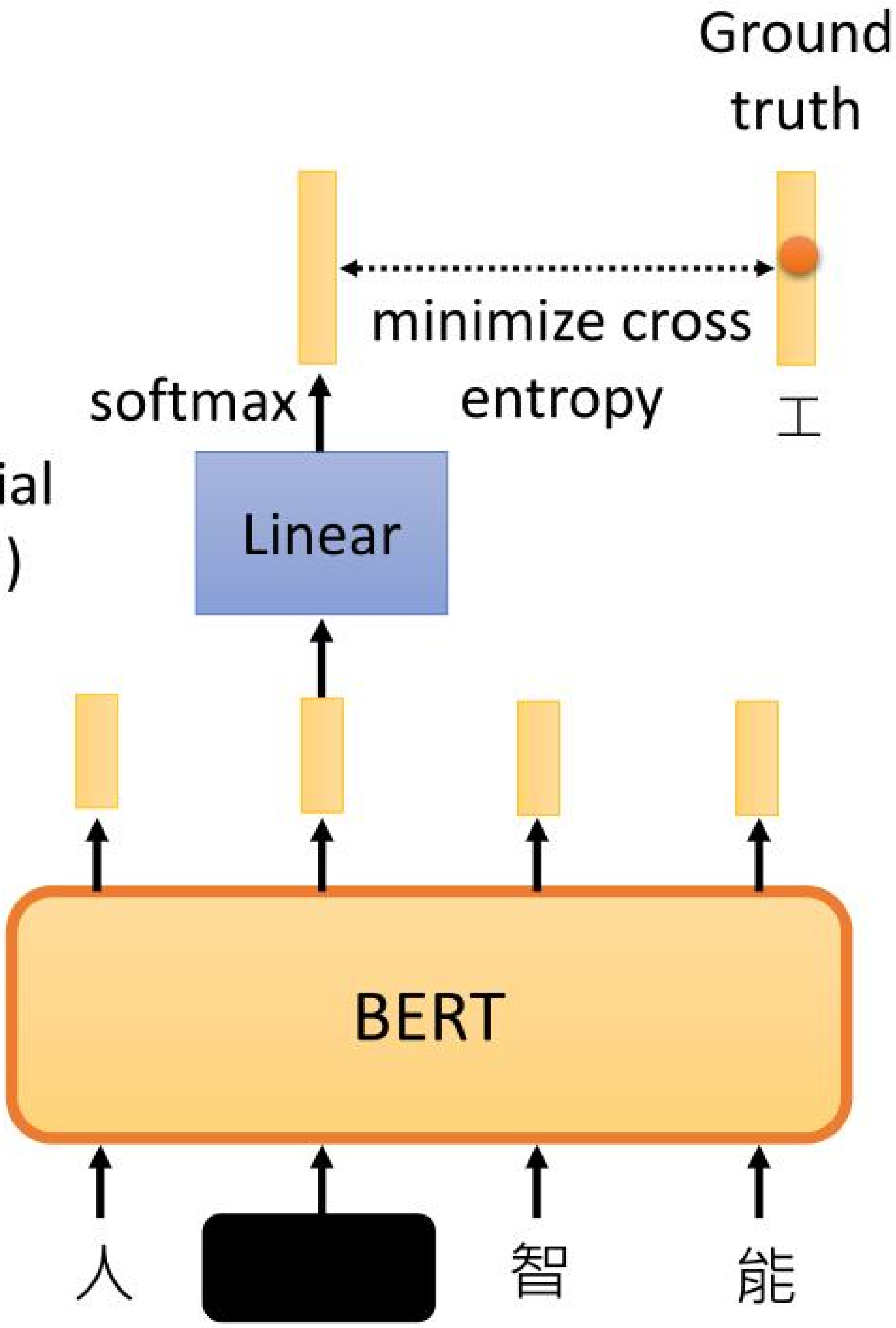
人	0.1
工	0.7
智	0.1
能	0.1
……	……

Randomly masking some tokens

Masking Input

https://arxiv.org/abs/1810.04805

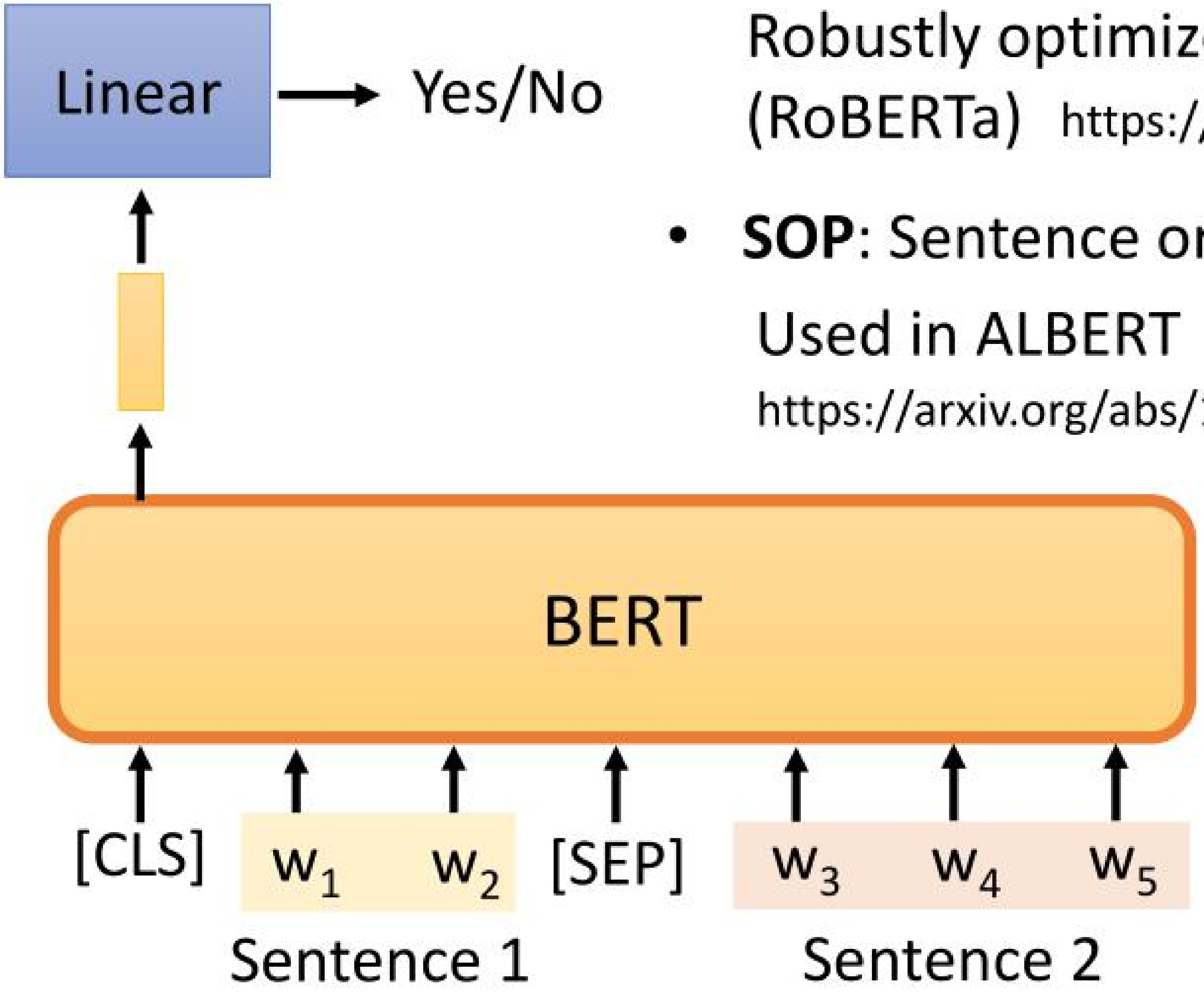




Transformer Encoder

Randomly masking some tokens

Next Sentence Prediction

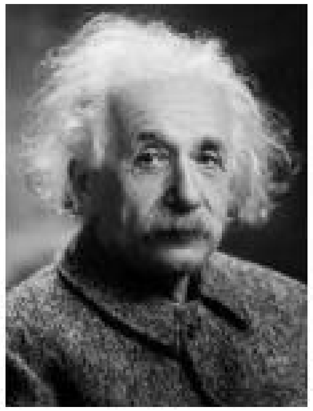


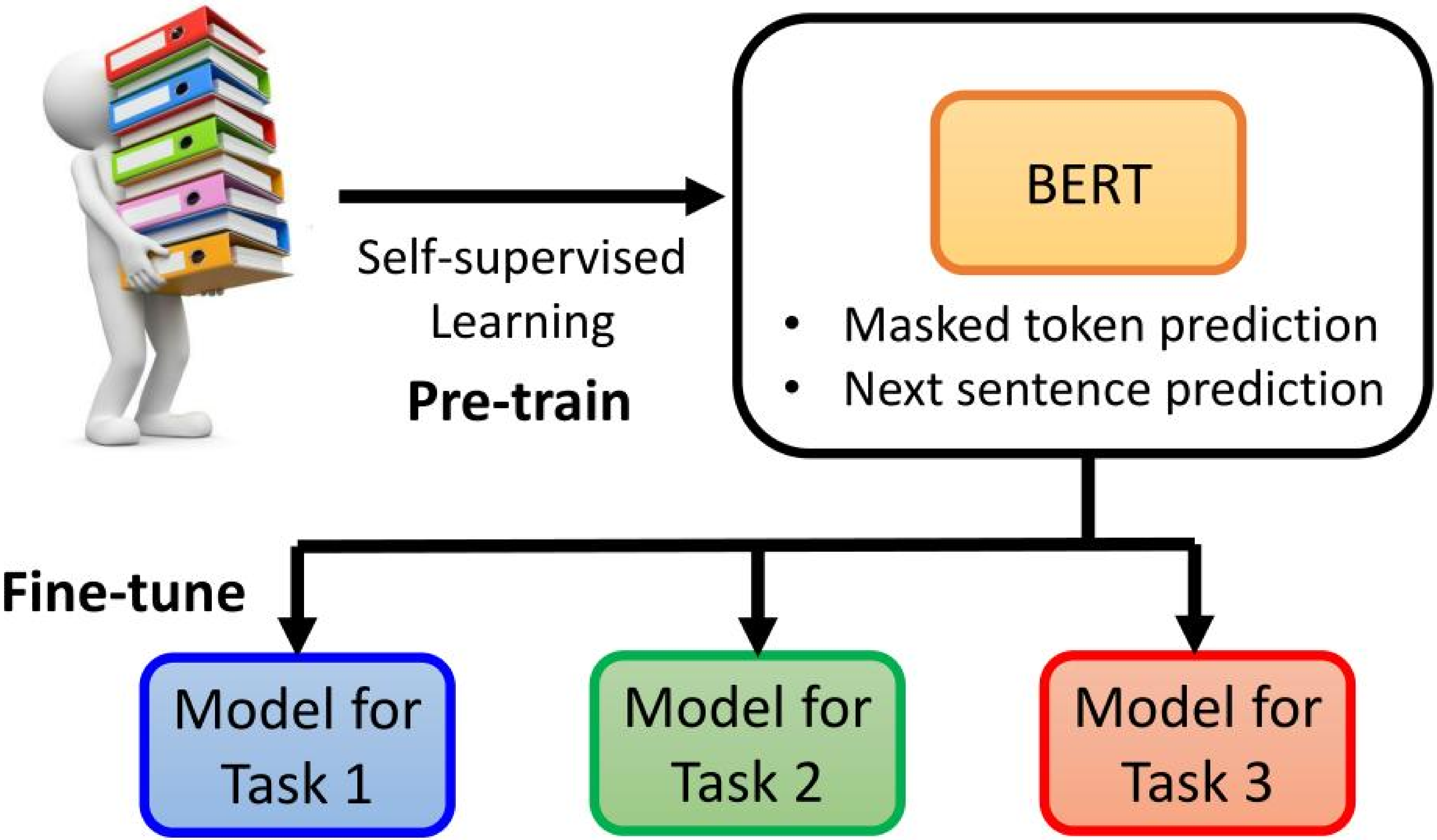
- This approach is not helpful.

  Robustly optimized BERT approach (RoBERTa) https://arxiv.org/abs/1907.11692

- **SOP**: Sentence order prediction

  Used in ALBERT https://arxiv.org/abs/1909.11942





Self-supervised Learning

Pre-train

BERT

- Masked token prediction
- Next sentence prediction

Fine-tune

Model for Task 1

Model for Task 2

Model for Task 3

Downstream Tasks

- The tasks we care
- We have a little bit labeled data.

GLUE

General Language Understanding Evaluation (GLUE)

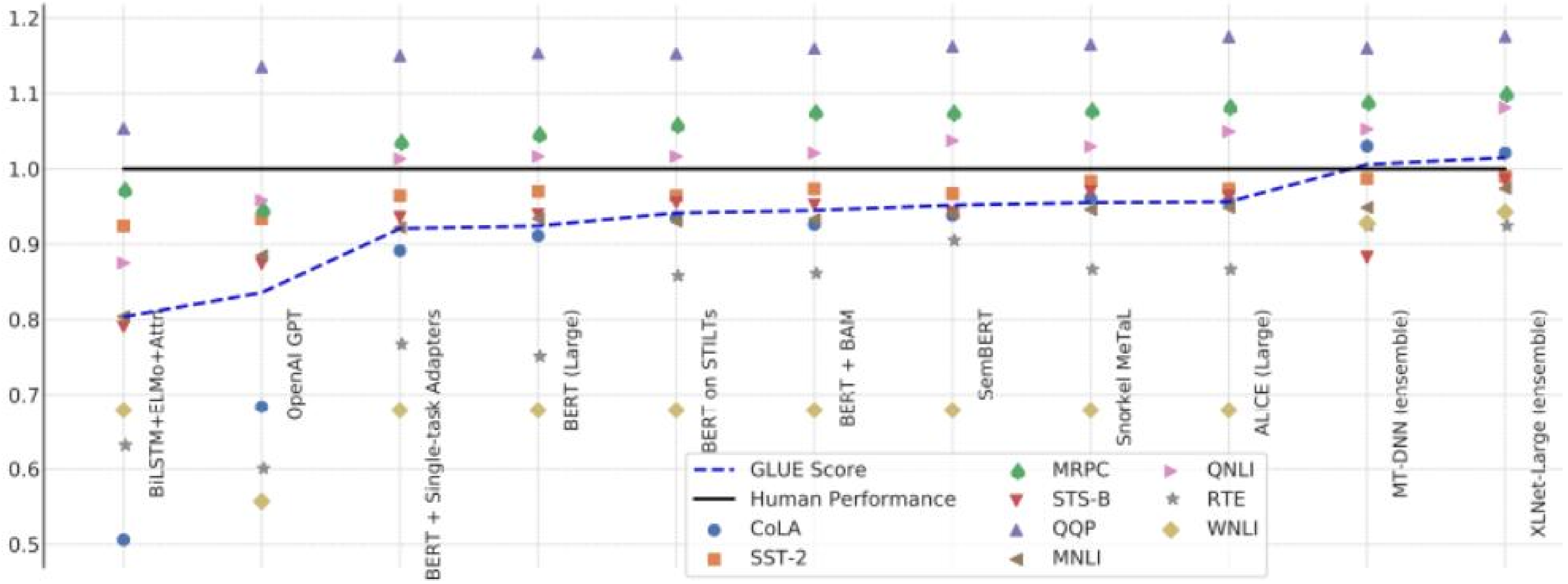
https://gluebenchmark.com/

- Corpus of Linguistic Acceptability (CoLA)
- Stanford Sentiment Treebank (SST-2)
- Microsoft Research Paraphrase Corpus (MRPC)
- Quora Question Pairs (QQP)
- Semantic Textual Similarity Benchmark (STS-B)
- Multi-Genre Natural Language Inference (MNLI)
- Question-answering NLI (QNLI)
- Recognizing Textual Entailment (RTE)
- Winograd NLI (WNLI)

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

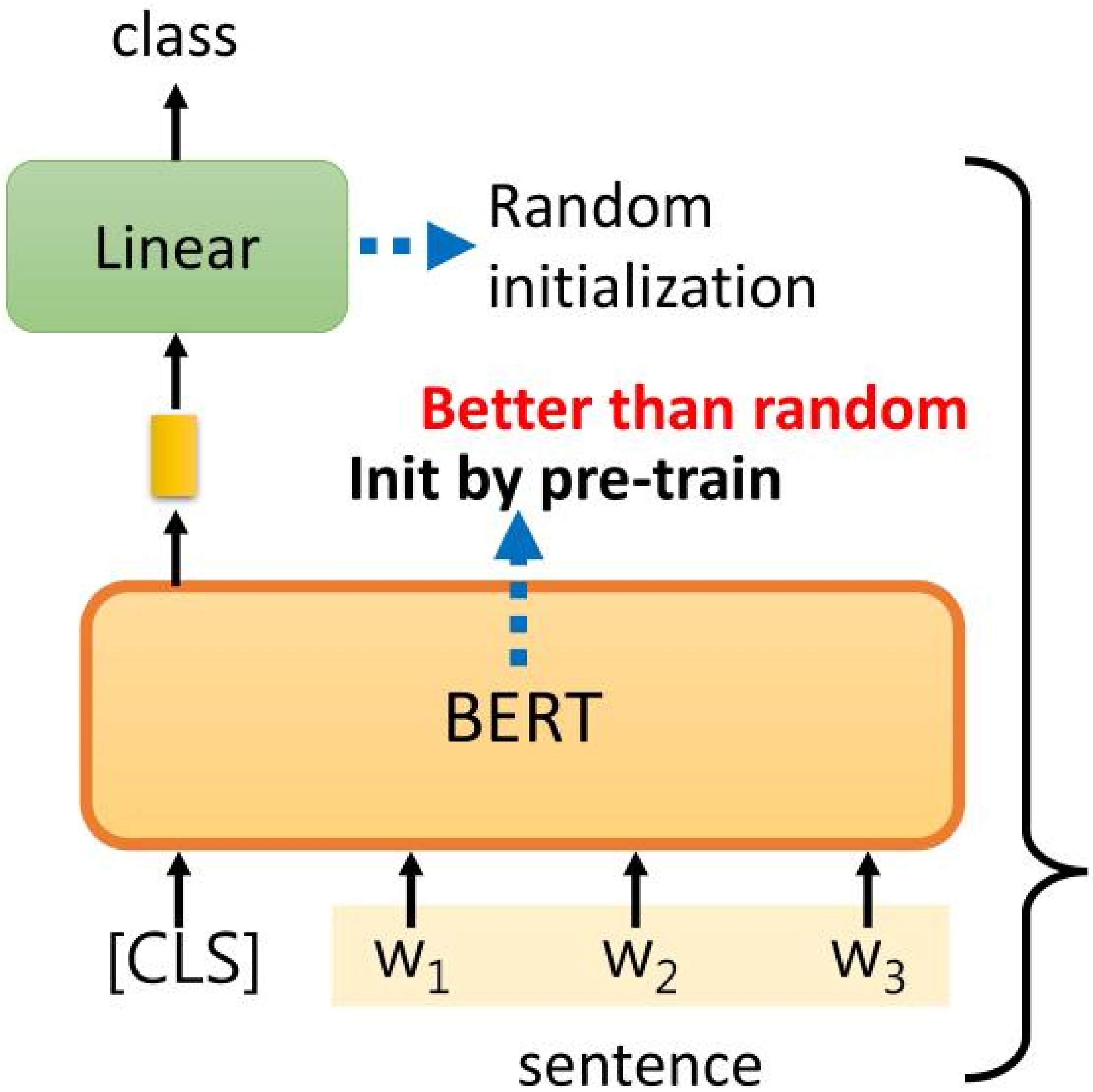
BERT and its Family

- GLUE scores



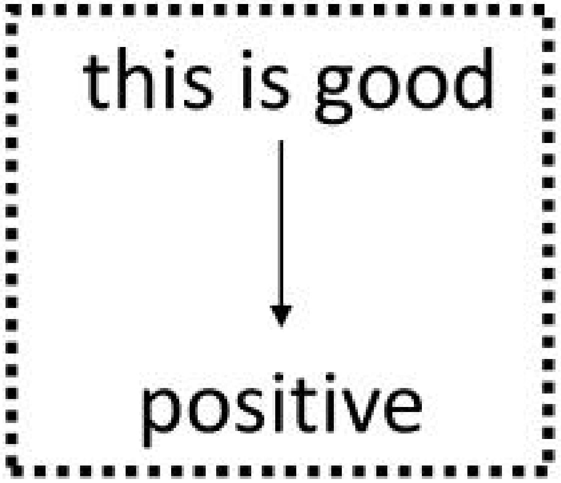
Source of image: https://arxiv.org/abs/1905.00537

How to use BERT – Case 1



Input: sequence
output: class

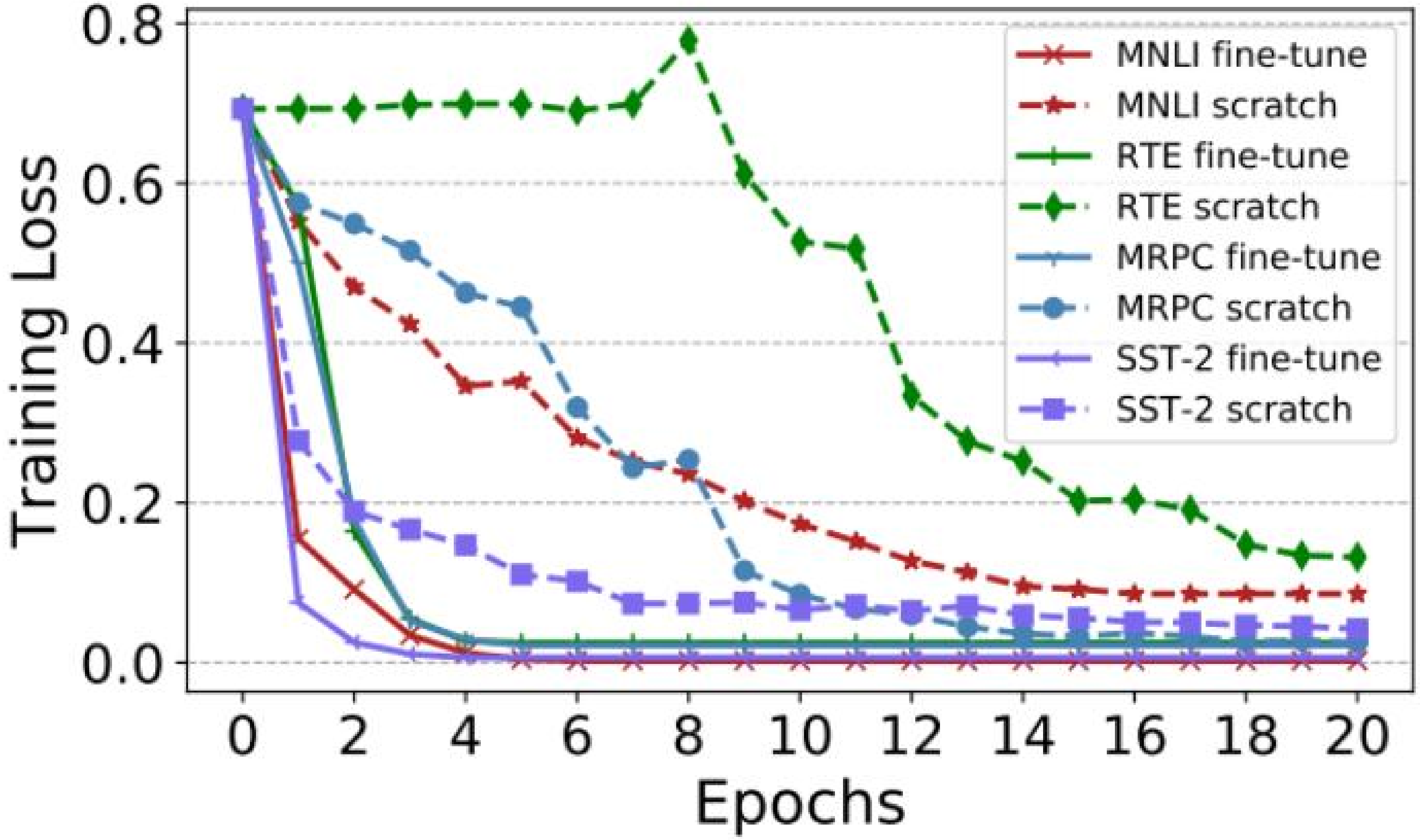
Example:
Sentiment analysis



This is the model to be learned.

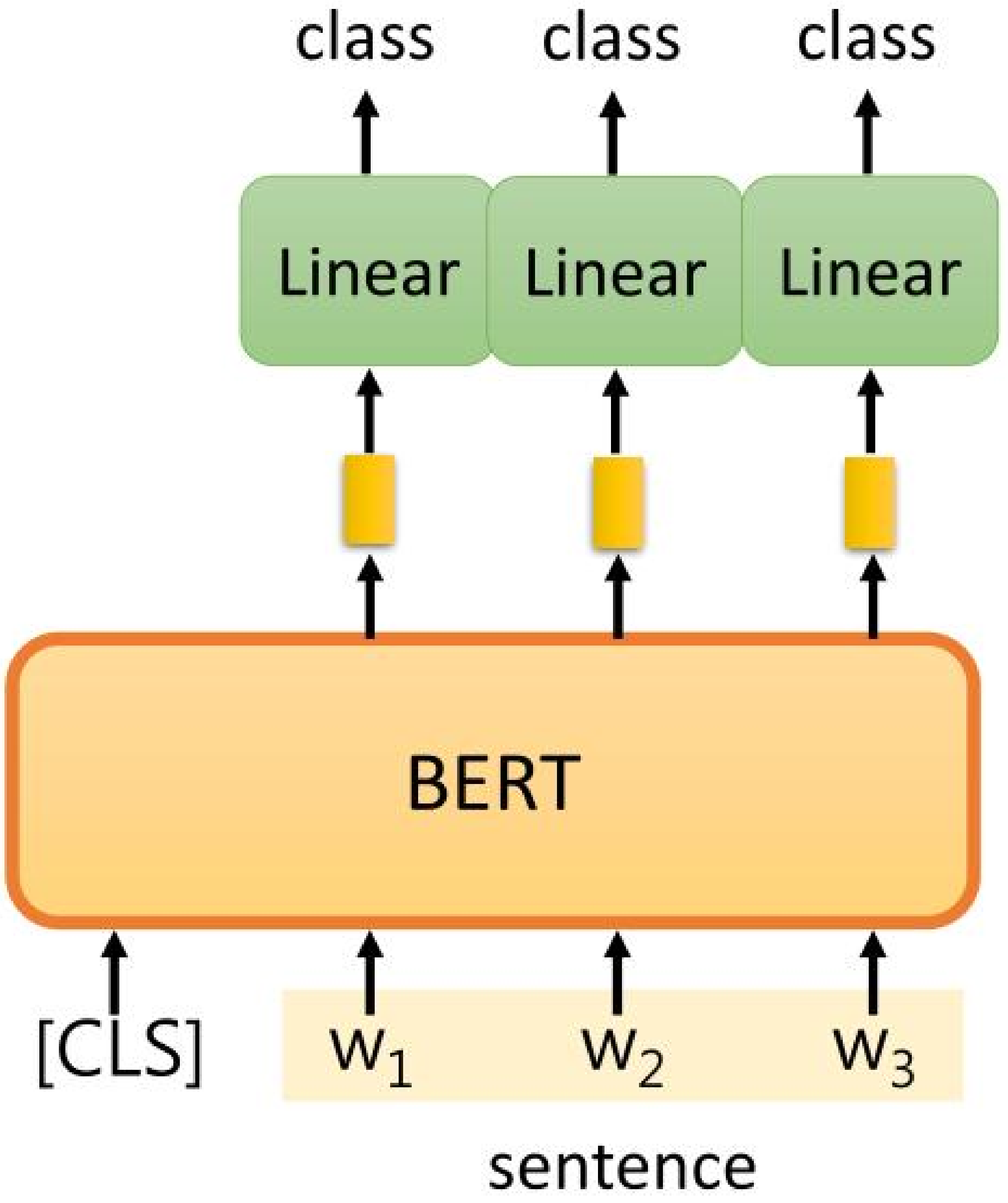
Pre-train v.s. Random Initialization

(fine-tune) (scratch)



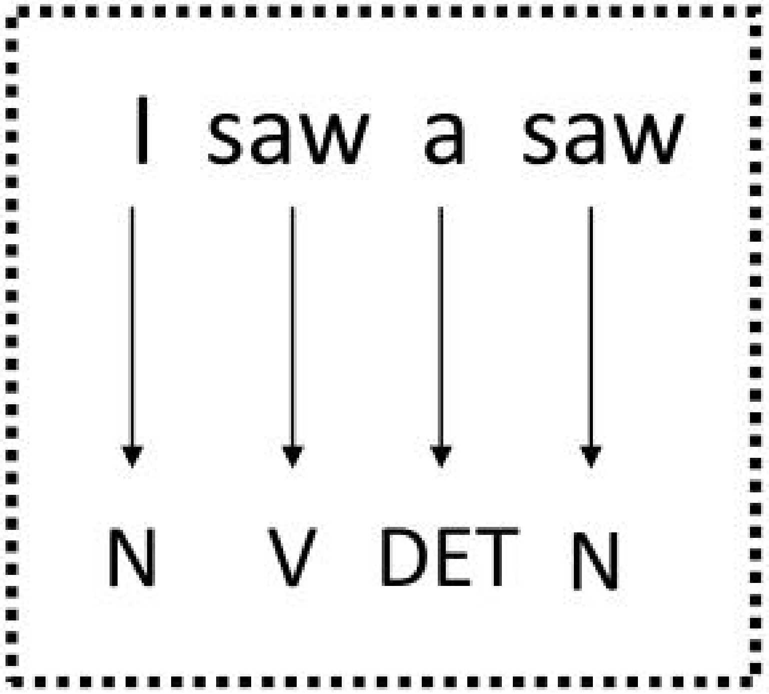
Source of image: https://arxiv.org/abs/1908.05620

How to use BERT – Case 2



Input: sequence
output: same as input

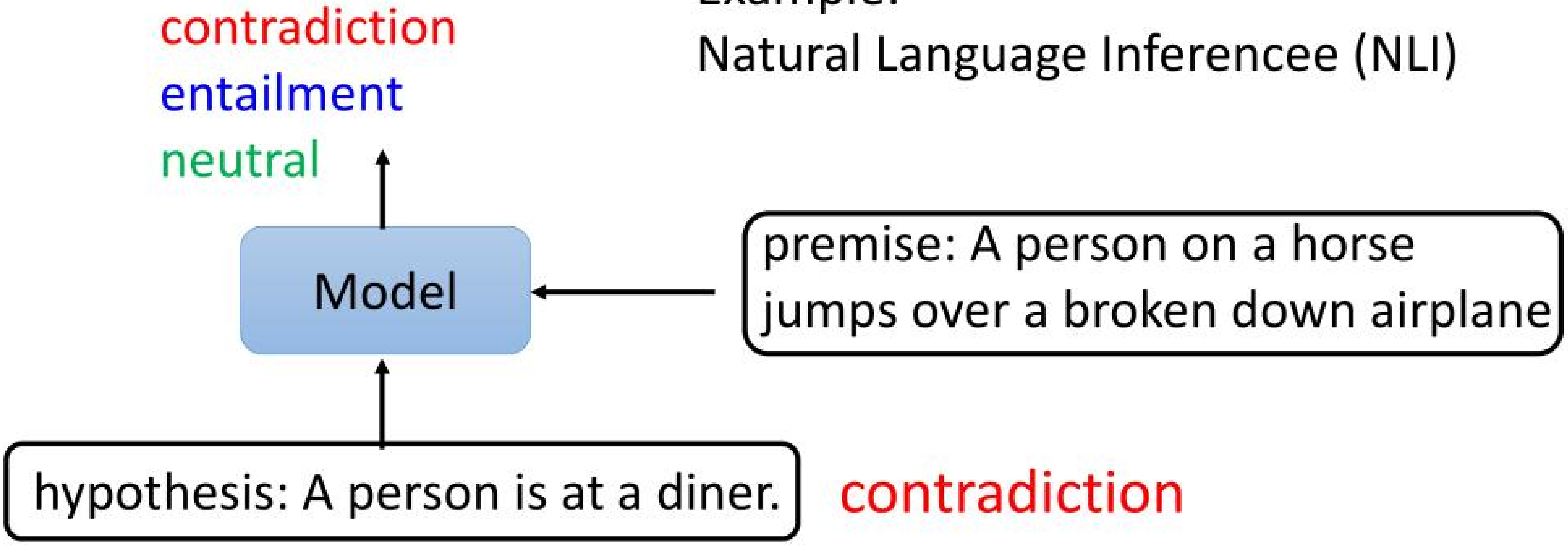
Example:
POS tagging



How to use BERT – Case 3

Input: two sequences
Output: a class

Example:
Natural Language Inferencee (NLI)



contradiction
entailment
neutral

Model

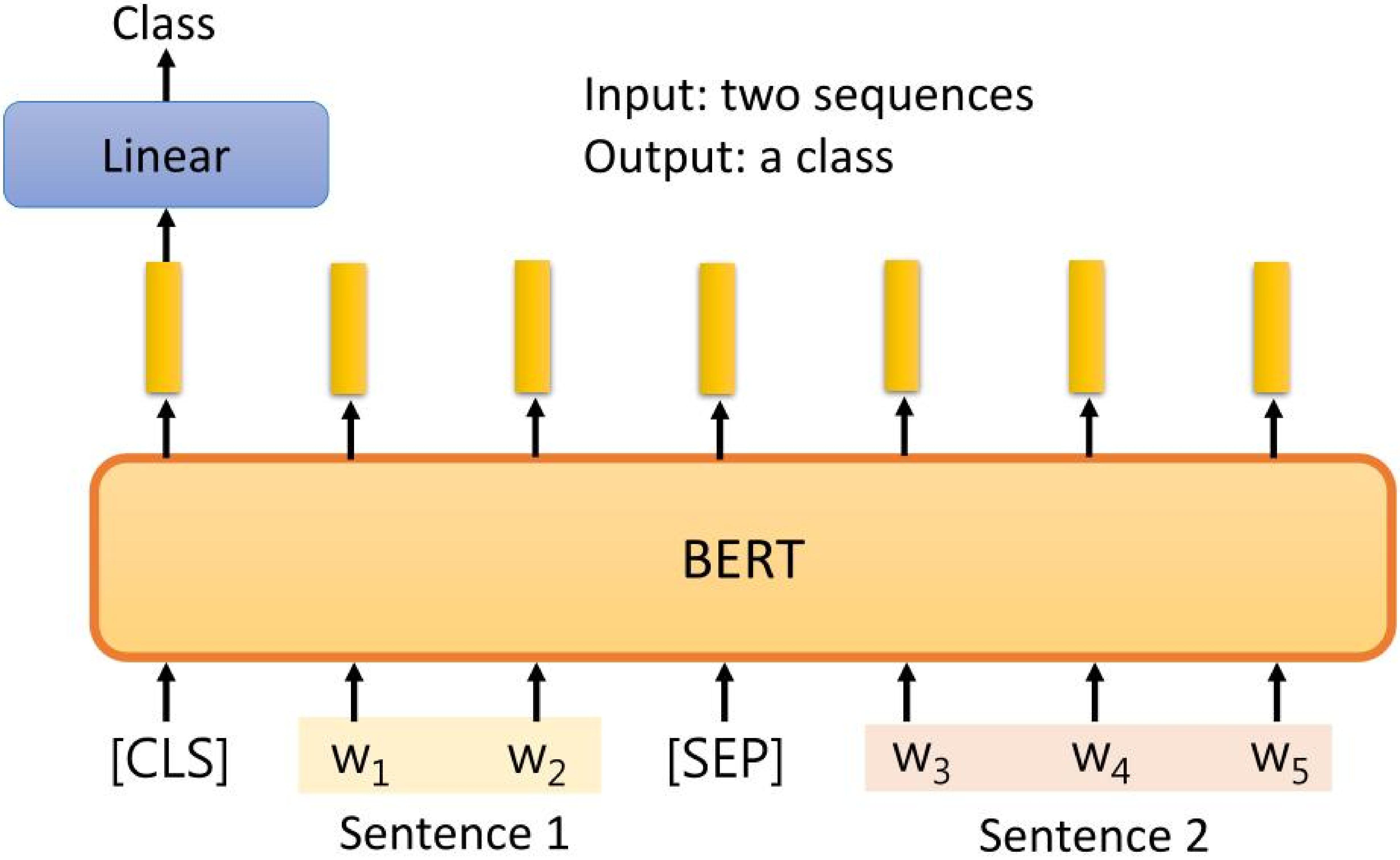
premise: A person on a horse jumps over a broken down airplane

hypothesis: A person is at a diner.

contradiction

How to use BERT – Case 3

Input: two sequences
Output: a class

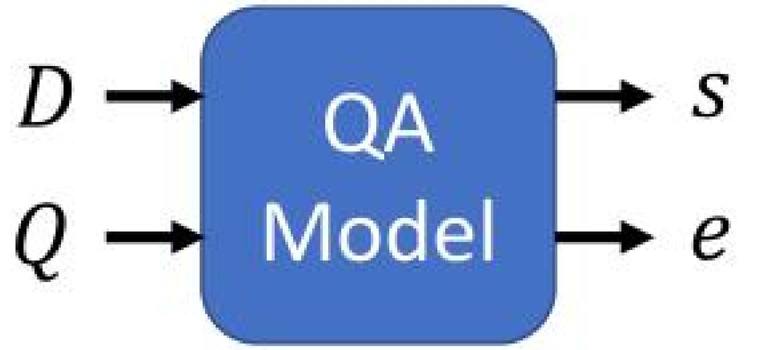


How to use BERT – Case 4

- Extraction-based Question Answering (QA)

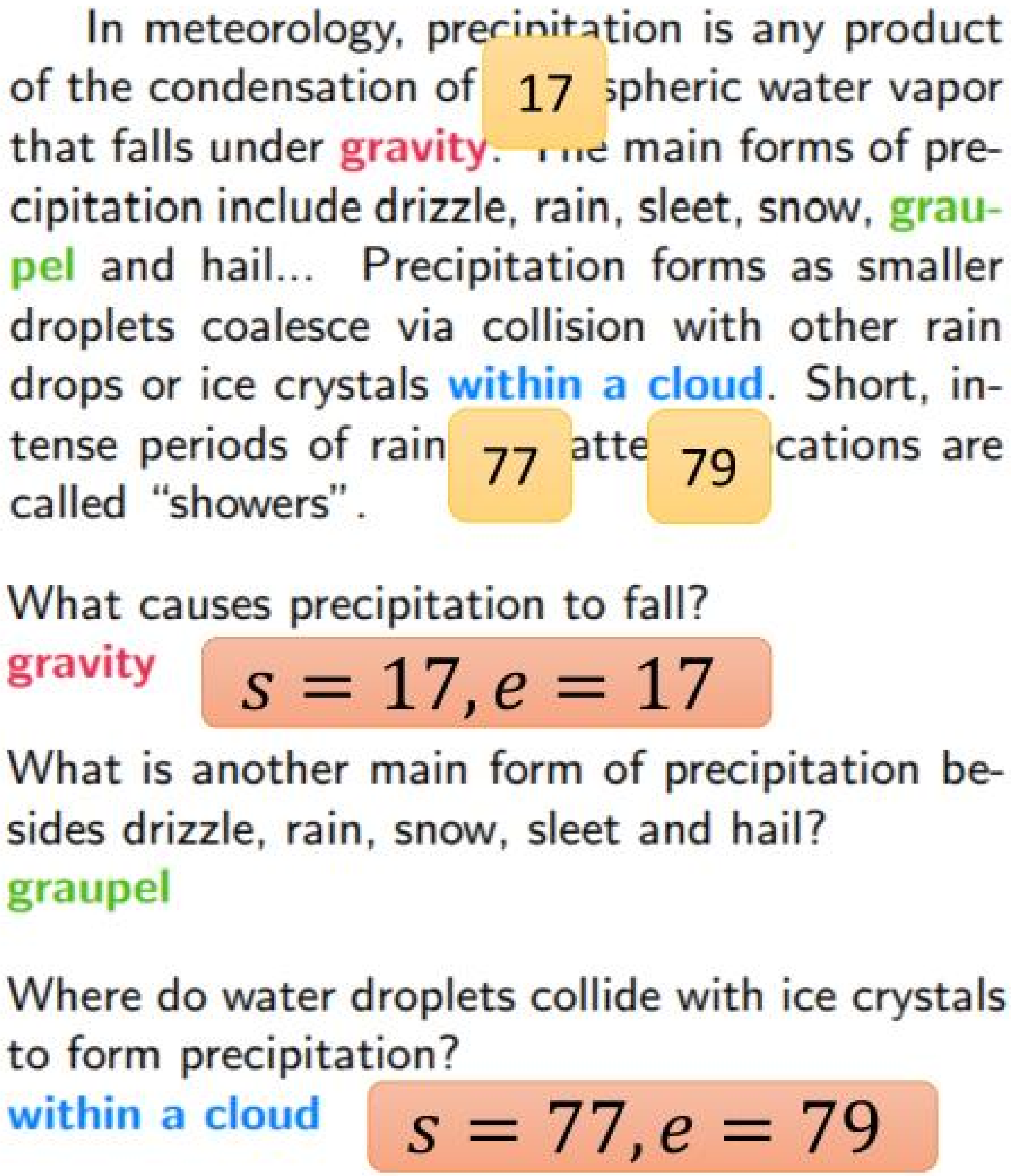
Document: $D = \{d_1, d_2, \cdots, d_N\}$

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 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?

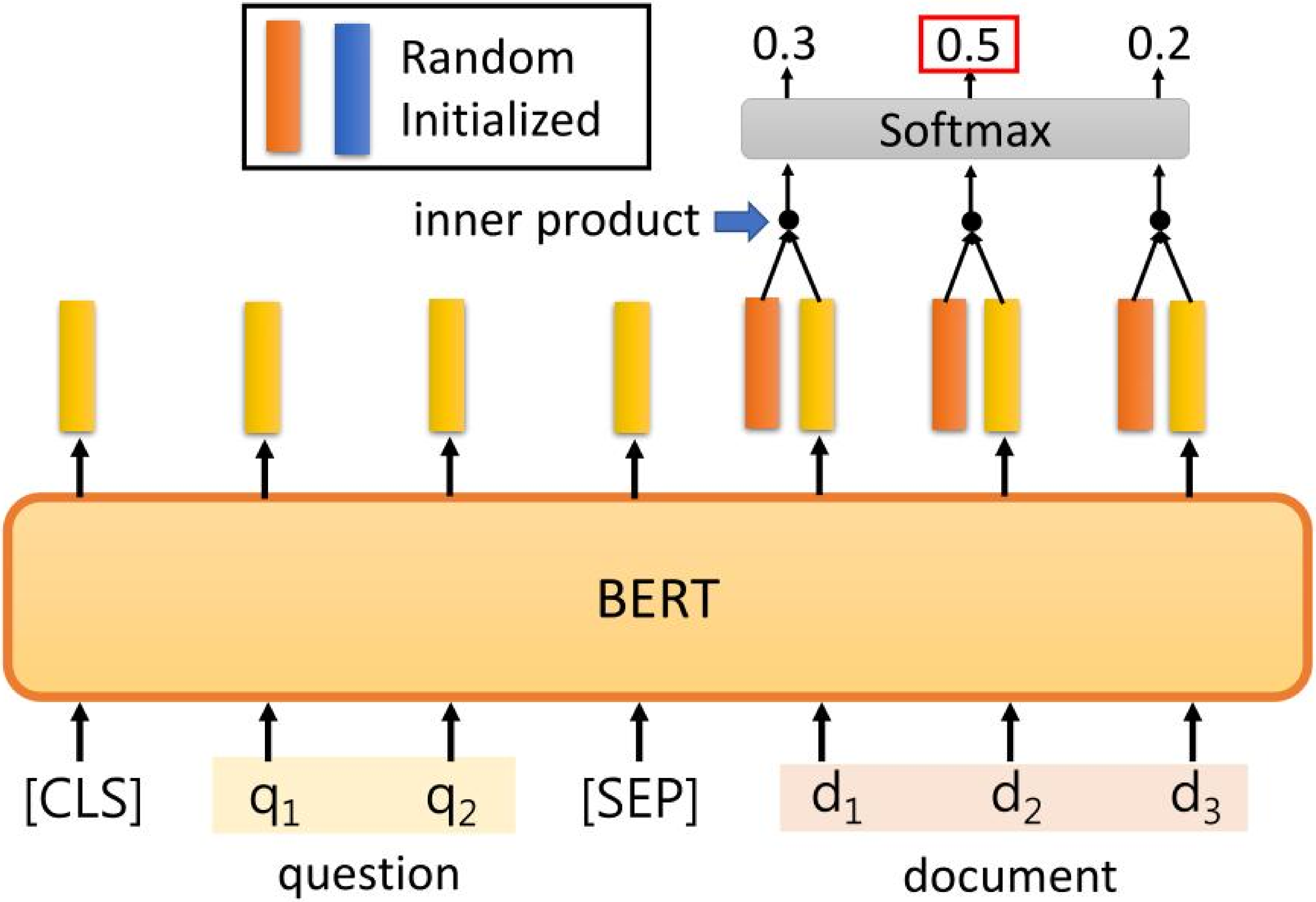
graupel

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

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

How to use BERT – Case 4

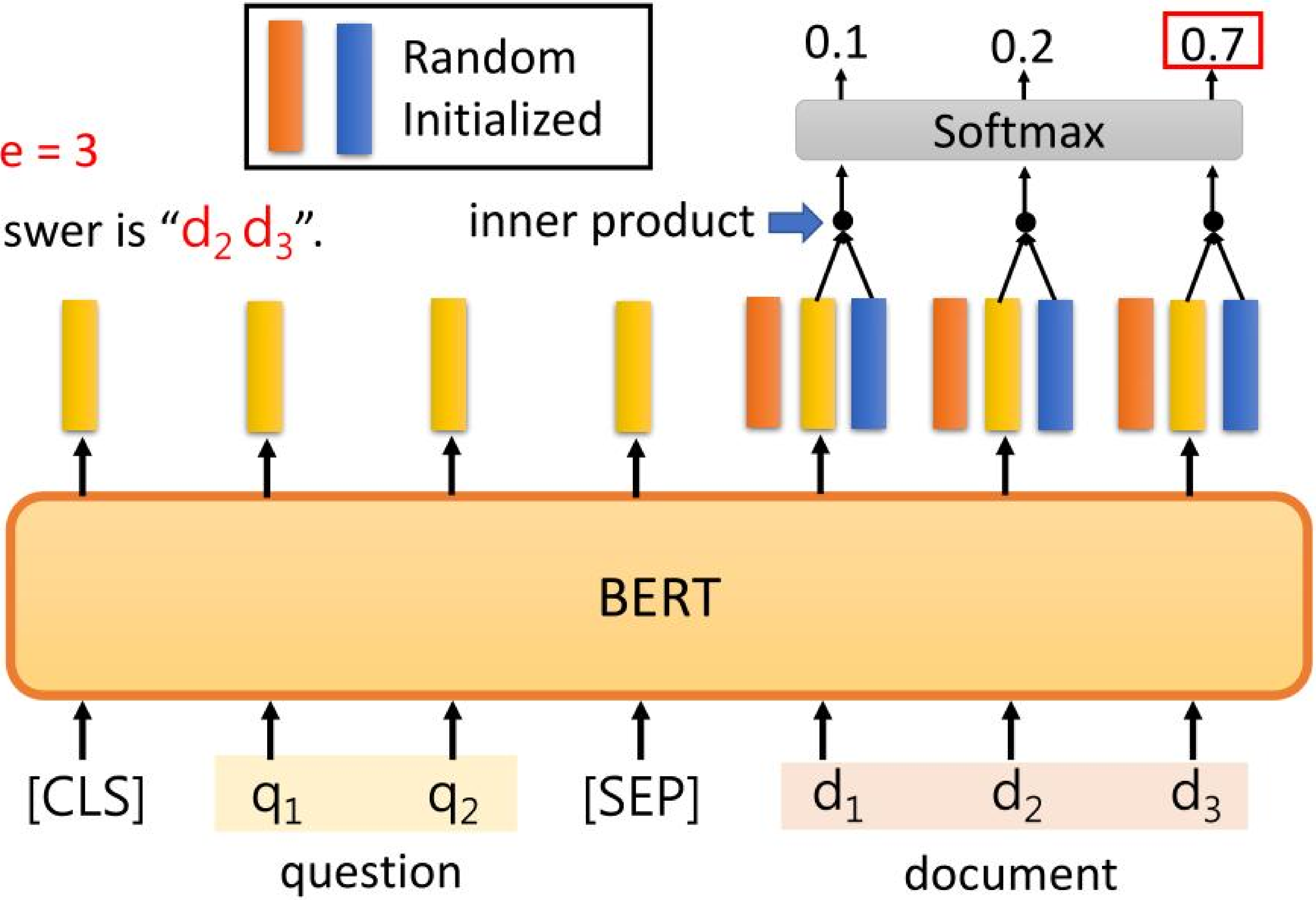
s = 2



How to use BERT – Case 4

s = 2   e = 3

The answer is "d_2 d_3".



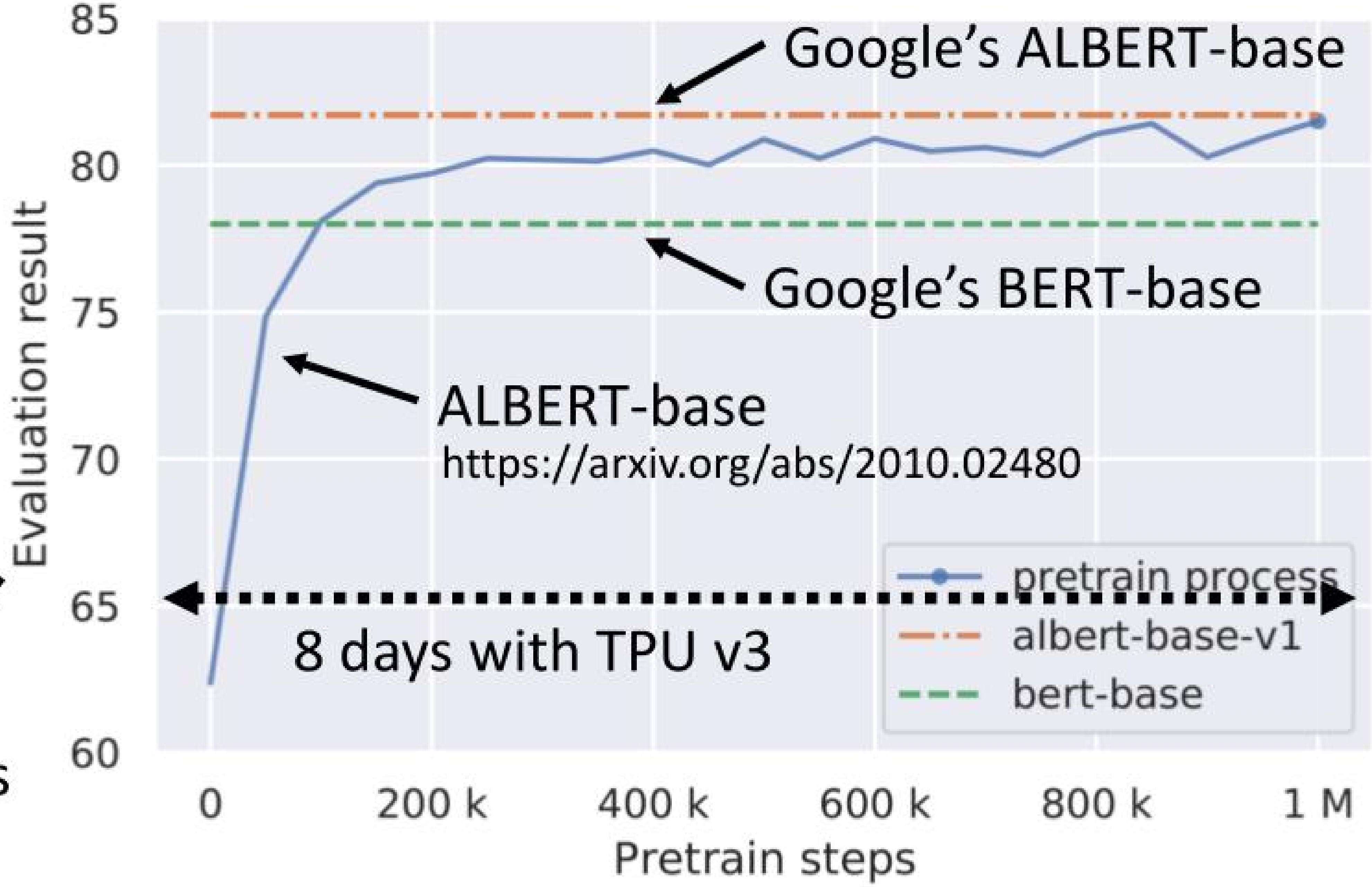
That's all!

designed by freepik.com

Training BERT is challenging!

Training data has more than **3 billions** of words.

3000 times of **Harry Potter series**



Google's ALBERT-base

Google's BERT-base

ALBERT-base

https://arxiv.org/abs/2010.02480

8 days with TPU v3

GLUE scores

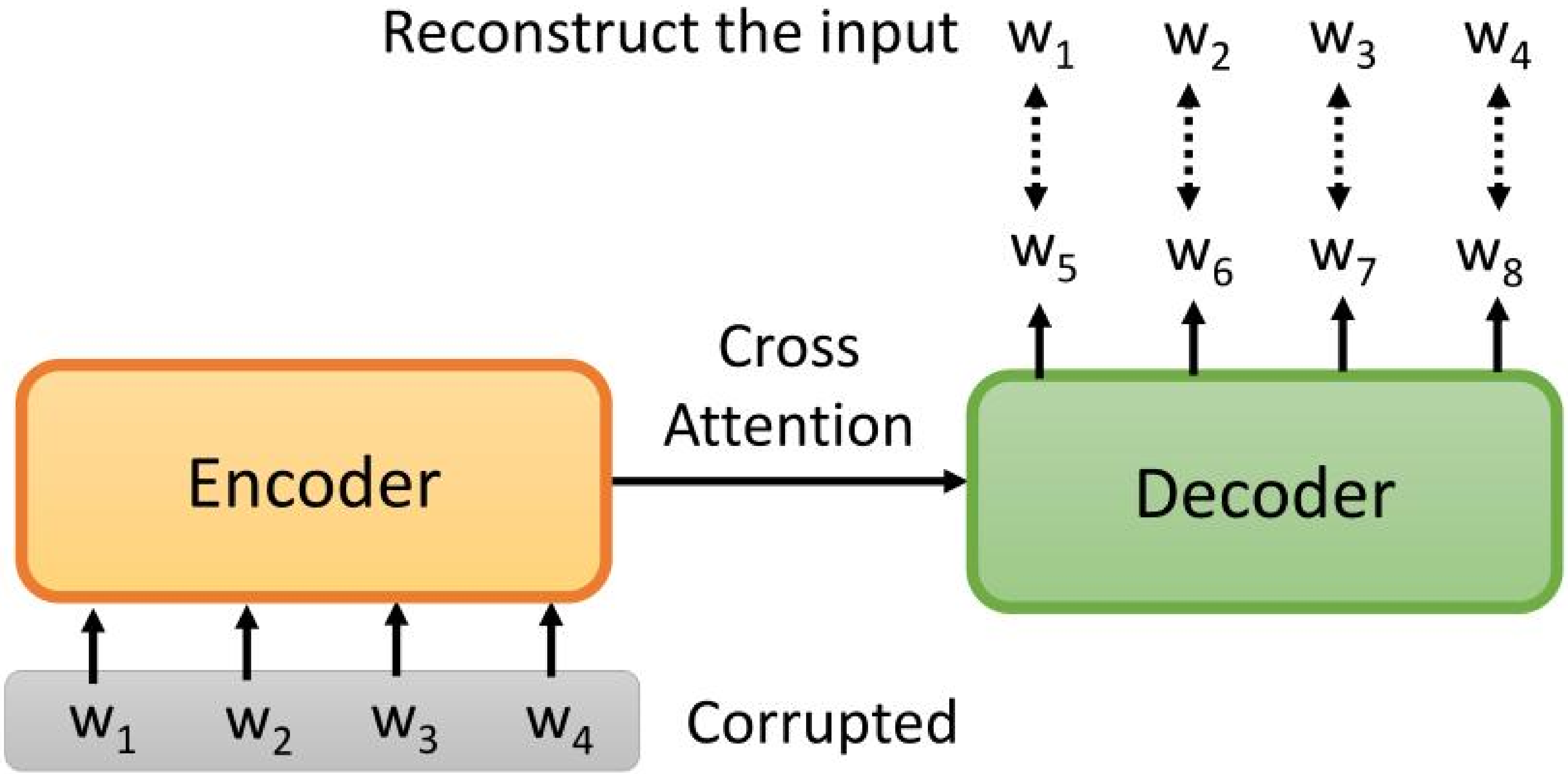
BERT Embryology (胚胎学)

https://arxiv.org/abs/2010.02480

When does BERT know POS tagging, syntactic parsing, semantics?

The answer is counterintuitive!

Pre-training a seq2seq model

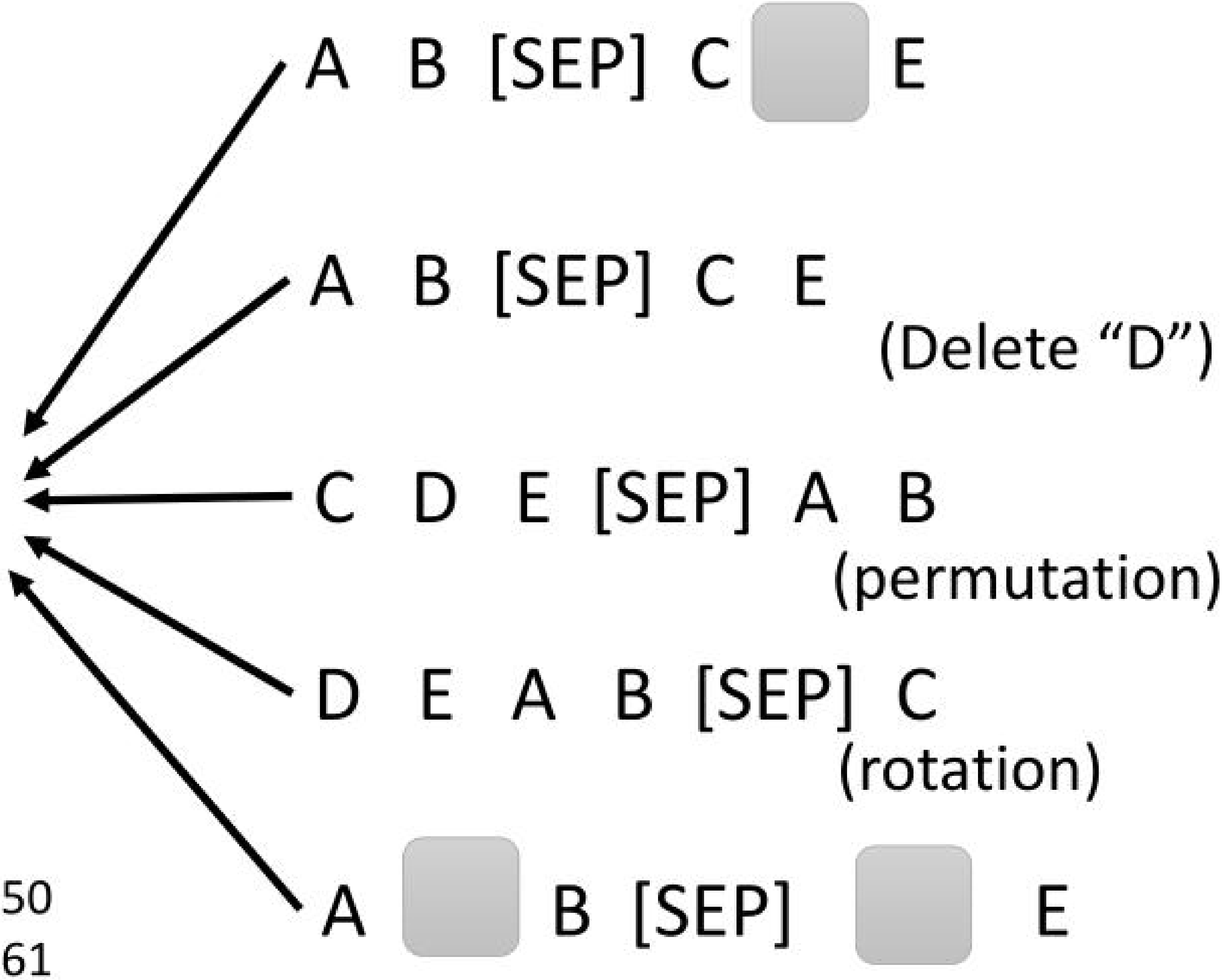


MASS / BART

BART

MASS

A B [SEP] C D E



A B [SEP] C ▢ E

A B [SEP] C E

(Delete "D")

C D E [SEP] A B

(permutation)

D E A B [SEP] C

(rotation)

A ▢ B [SEP] ▢ E

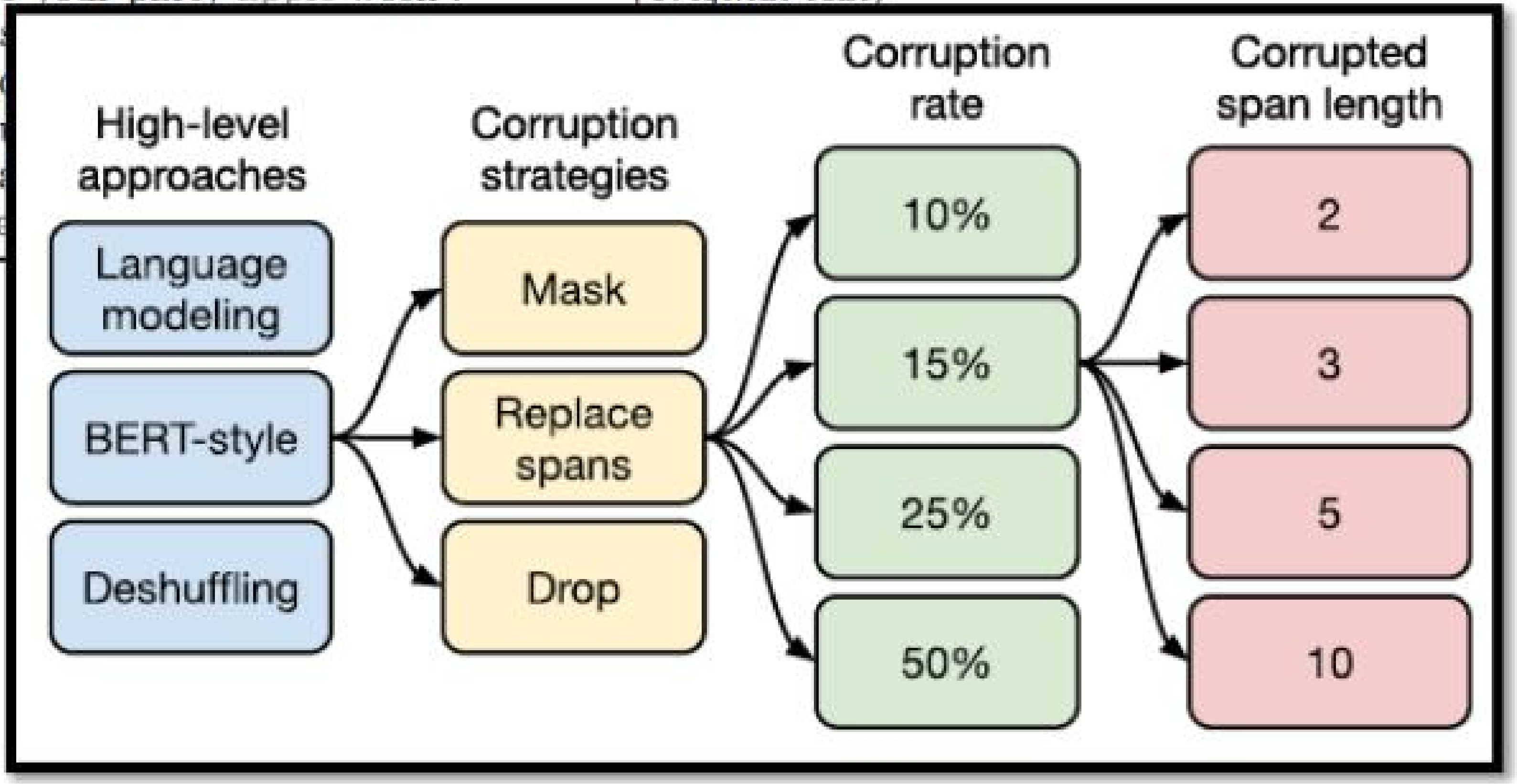
Text Infilling

https://arxiv.org/abs/1905.02450
https://arxiv.org/abs/1910.13461

T5 – Comparison

- Transfer Text-to-Text Transformer (T5)
- Colossal Clean Crawled Corpus (C4)

Objective	Inputs	Targets
Prefix language modeling	Thank you for inviting	me to your party last week .
BERT-style	Thank you <M> <M> me to your party apple week .	*(original text)*
Deshuffling	party me for your to . la	
I.i.d. noise, mask tokens	Thank you <M> <M> me t	
I.i.d. noise, replace spans	Thank you <X> me to yo	
I.i.d. noise, drop tokens	Thank you me to your pa	
Random spans	Thank you <X> to <Y> we	

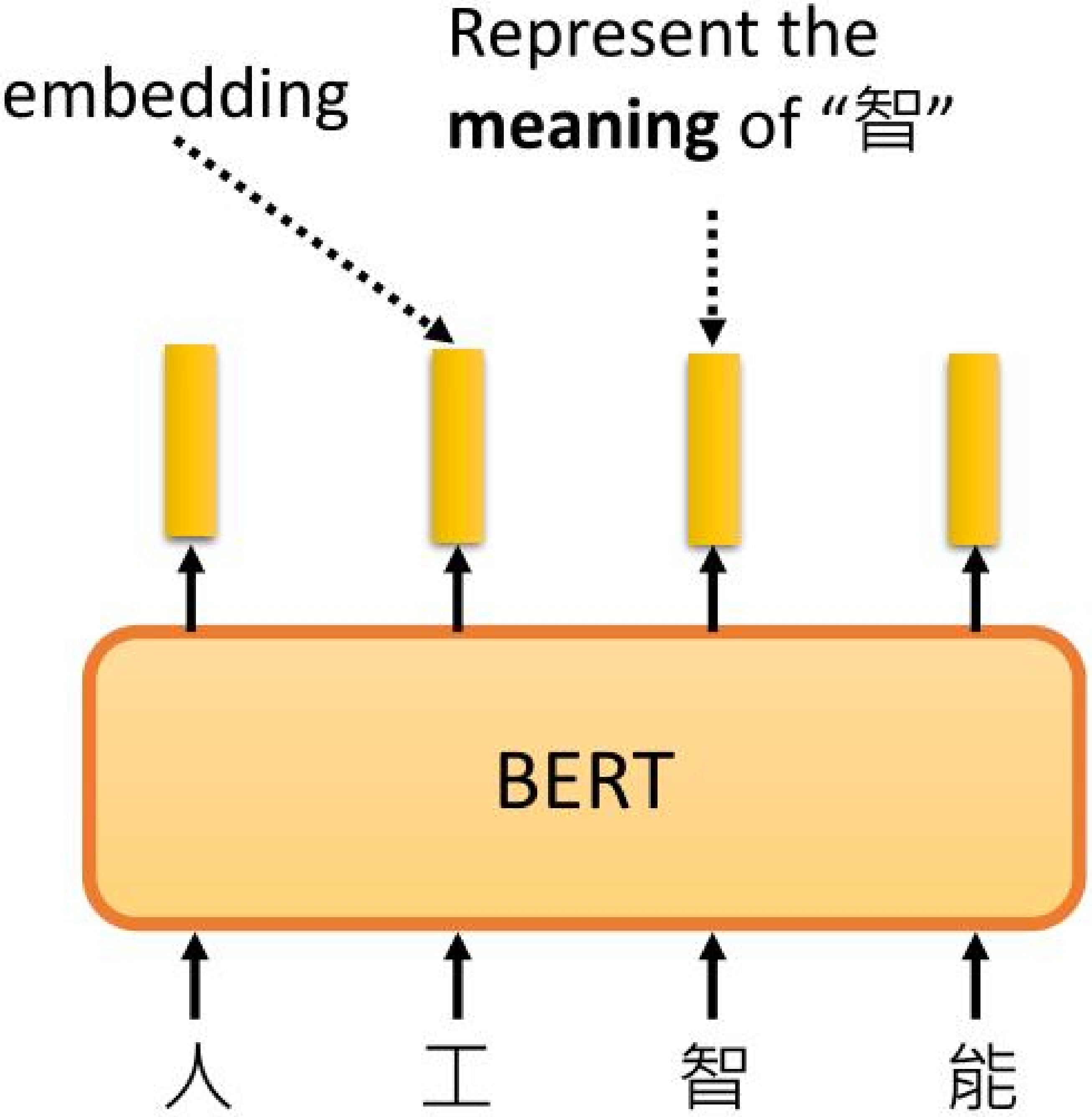

High-level approaches: Language modeling, BERT-style, Deshuffling

Corruption strategies: Mask, Replace spans, Drop

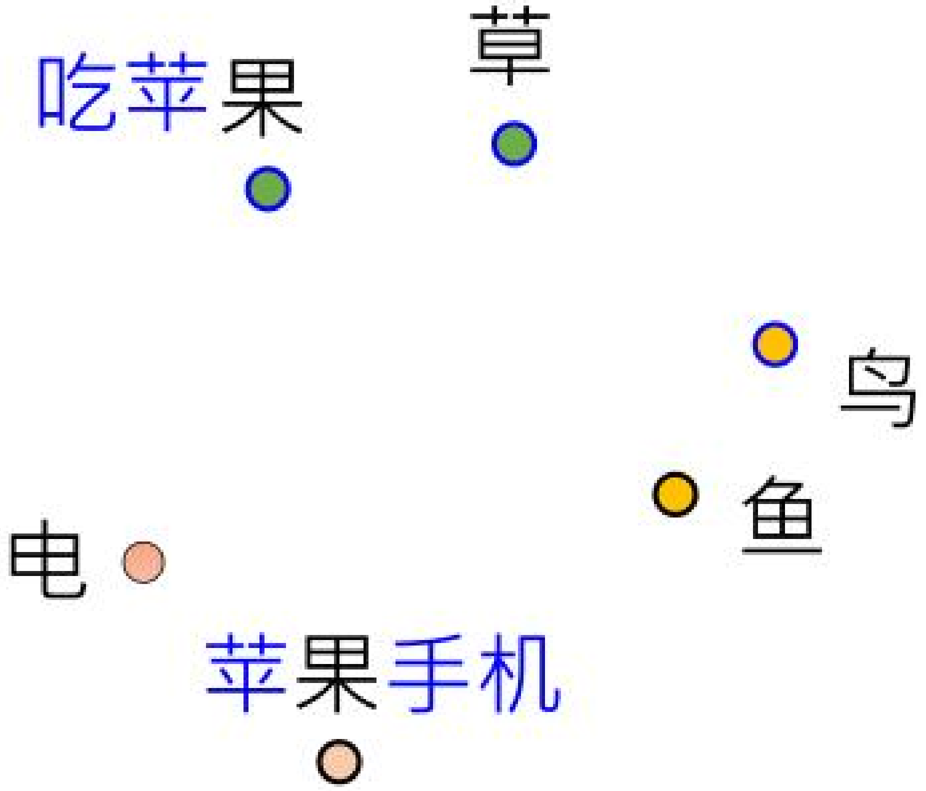
Corruption rate: 10%, 15%, 25%, 50%

Corrupted span length: 2, 3, 5, 10

Why does BERT work?

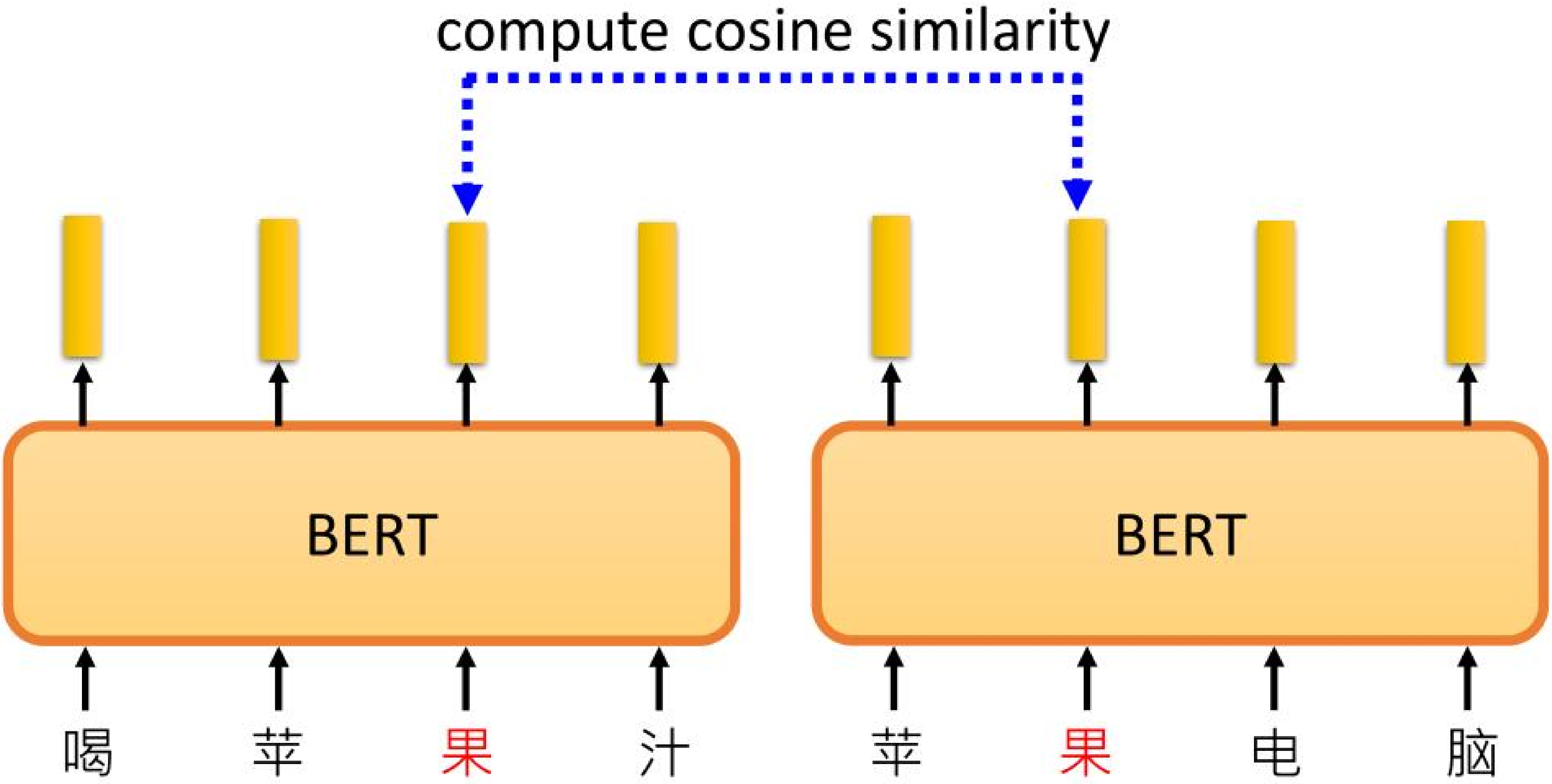


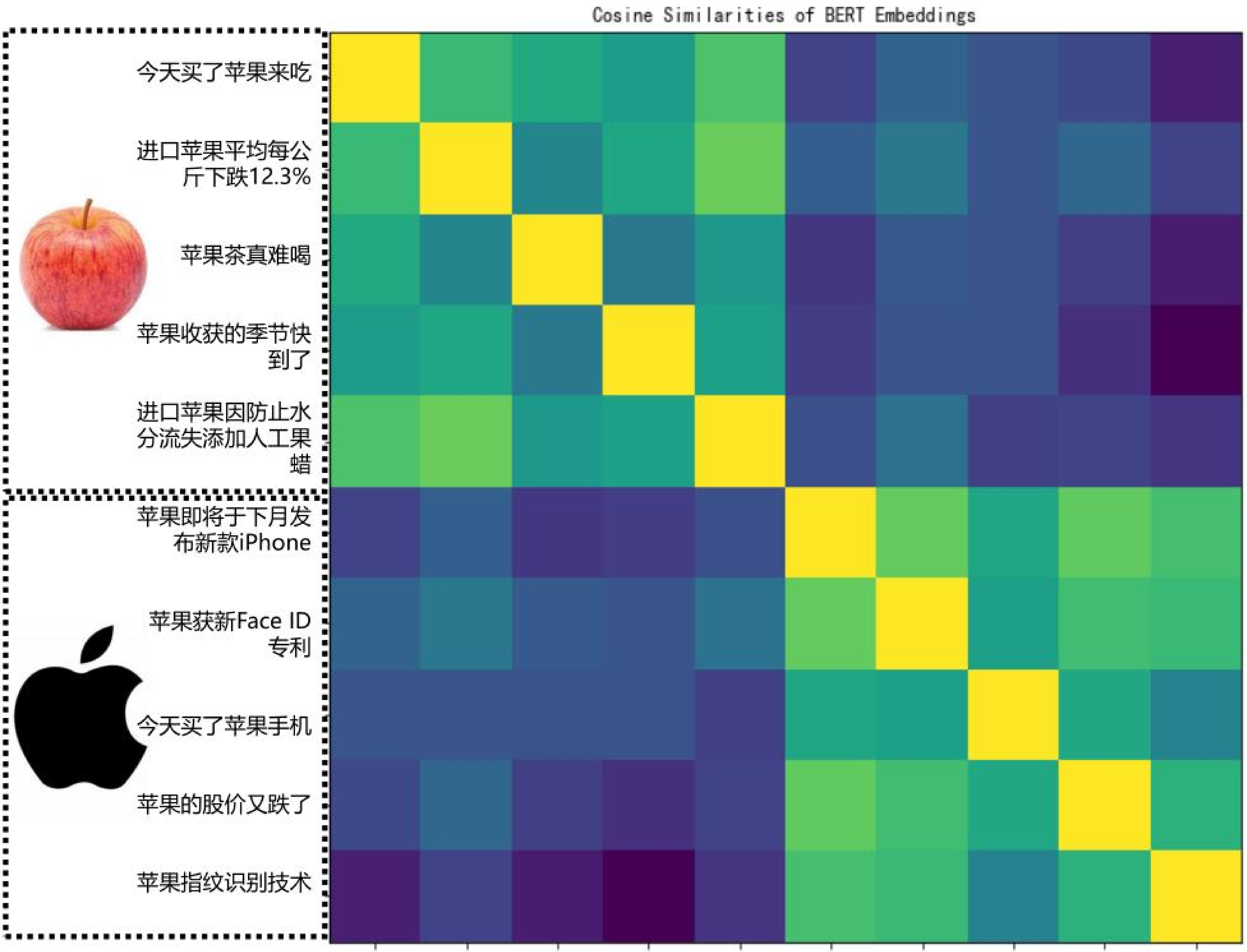
The tokens with similar meaning have similar embedding.



Context is considered.

Why does BERT work?



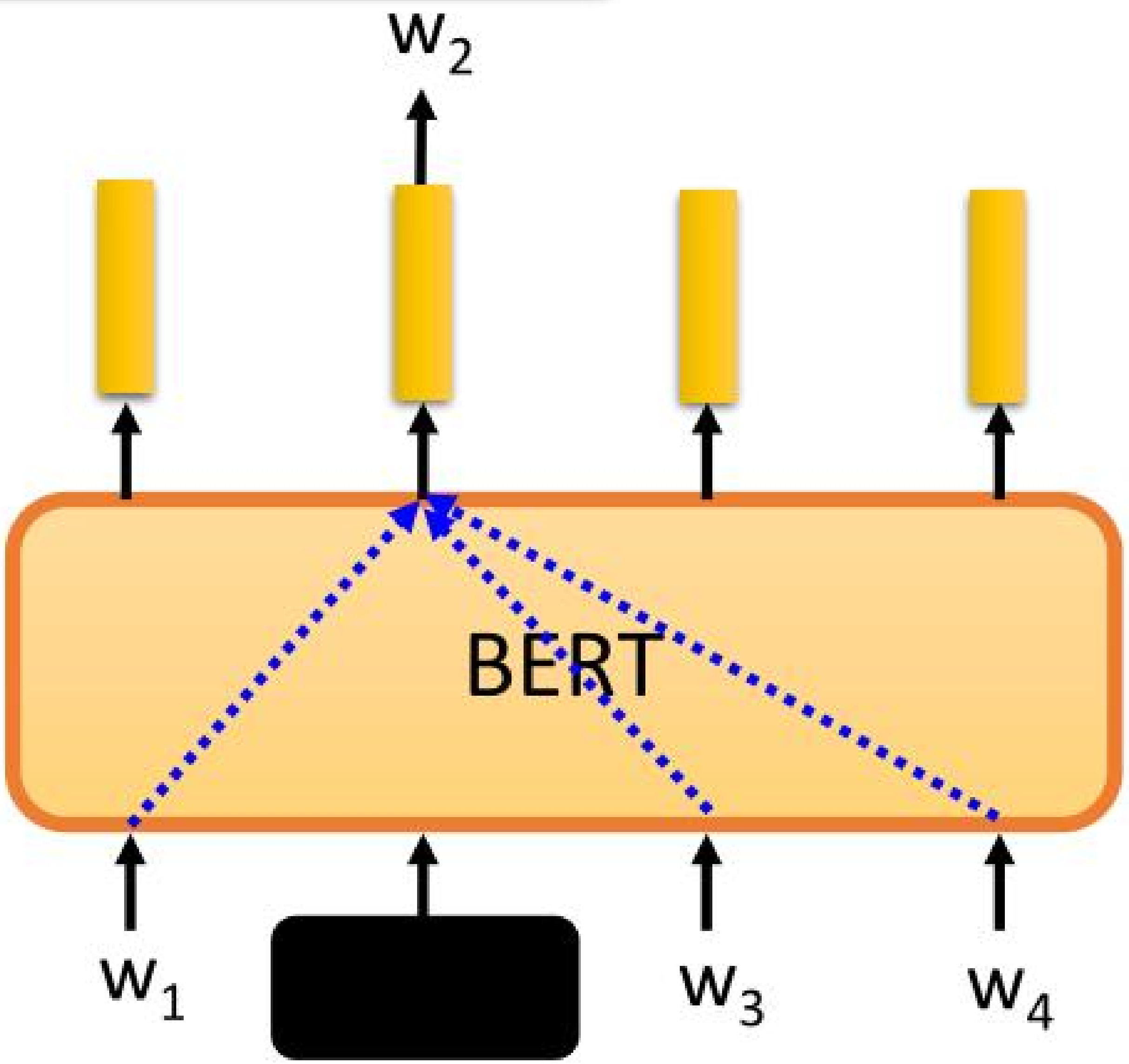

Cosine Similarities of BERT Embeddings
今天买了苹果来吃
进口苹果平均每公斤下跌12.3%
苹果茶真难喝
苹果收获的季节快到了
进口苹果因防止水分流失添加人工果蜡
苹果即将于下月发布新款iPhone
苹果获新Face ID专利
今天买了苹果手机
苹果的股价又跌了
苹果指纹识别技术

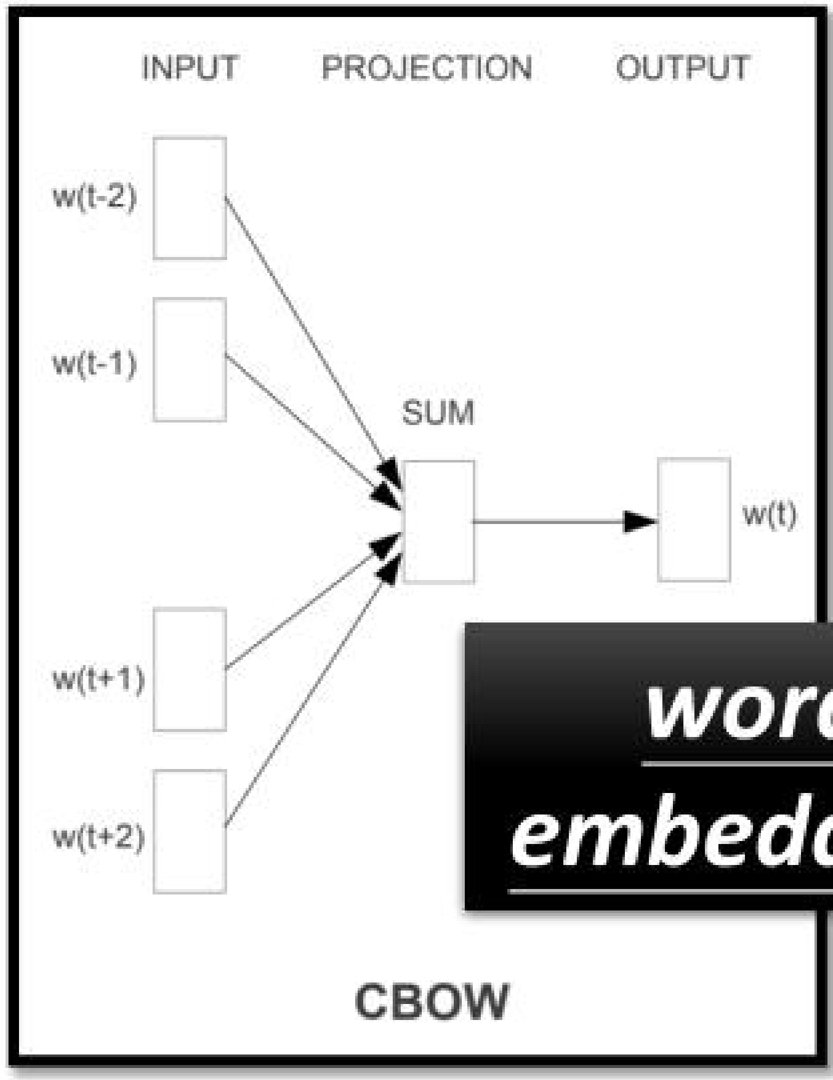
Why does BERT work?

Contextualized word embedding

You shall know a word by the company it keeps

John Rupert Firth





word embedding

https://arxiv.org/abs/2103.07162

Why does BERT work?

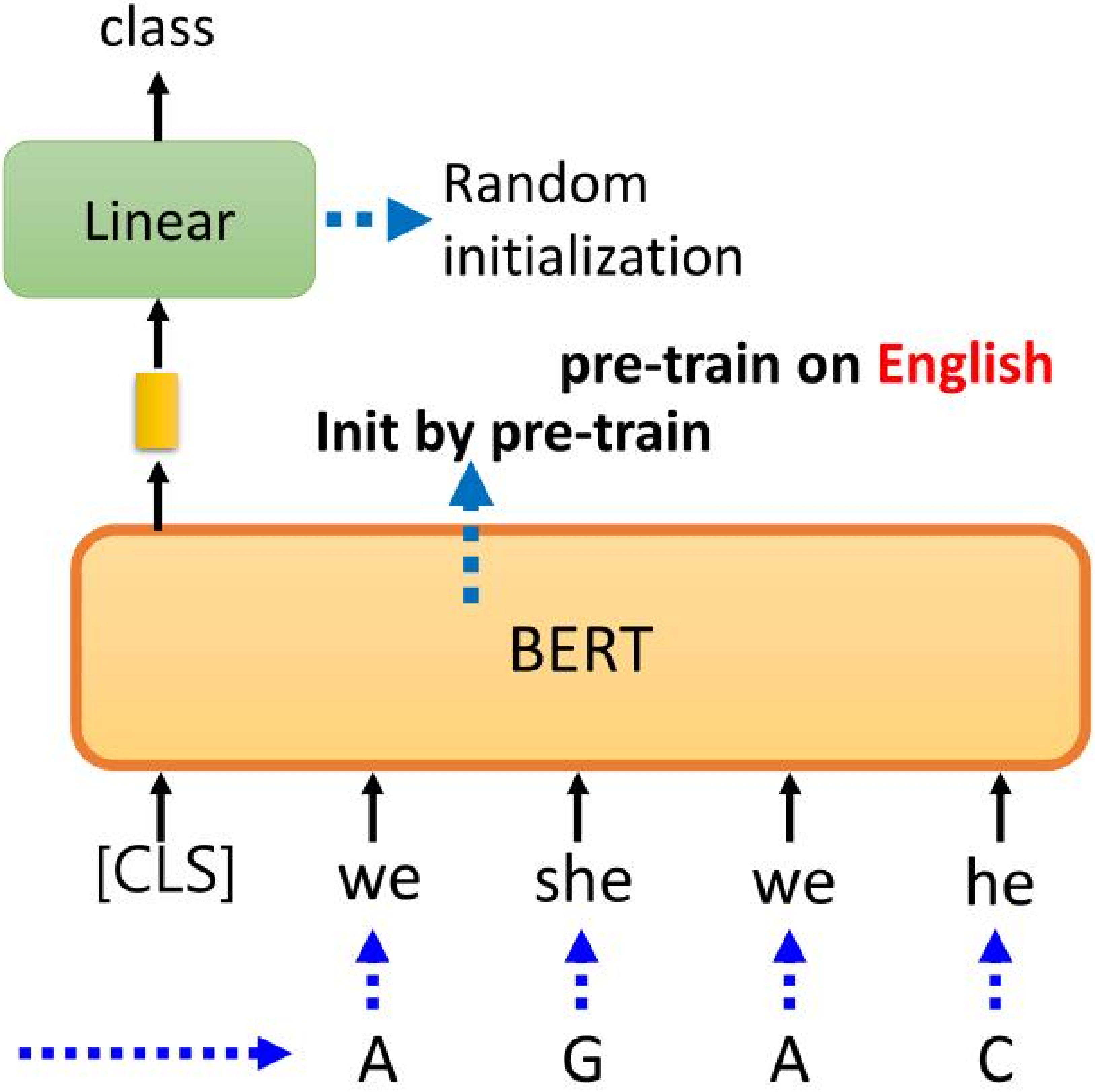
- Applying BERT to **protein**, **DNA**, **music classification**

EI	CCAGCTGCATCACAGGAGGCCAGCG
EI	AGACCCGCCGGGAGGCGGAGGACC
IE	AACGTGGCCTCCTTGTGCCCTTCCCC
IE	CCACTCAGCCAGGCCCTTCTTCTCCT
IE	CCTGATCTGGGTCTCCCCTCCCACCC
IE	AGCCCTCAACCCTTCTGTCTCACCCT
IE	CCACTCAGCCAGGCCCTTCTTCTCCT
N	CTGTGTTCACCACATCAAGCGCCGGG
N	GTGTTACCGAGGGCATTTCTAACAGT
N	TCTGAGCTCTGCATTTGTCTATTCTCC
class	DNA sequence

https://arxiv.org/abs/2103.07162

Why does BERT work?

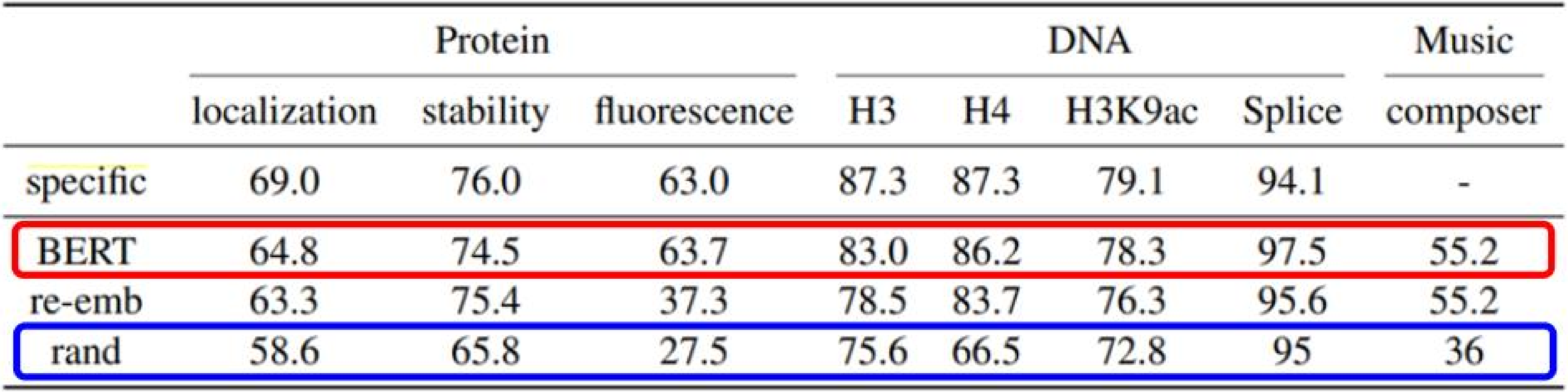
A	we
T	you
C	he
G	she



https://arxiv.org/abs/2103.07162

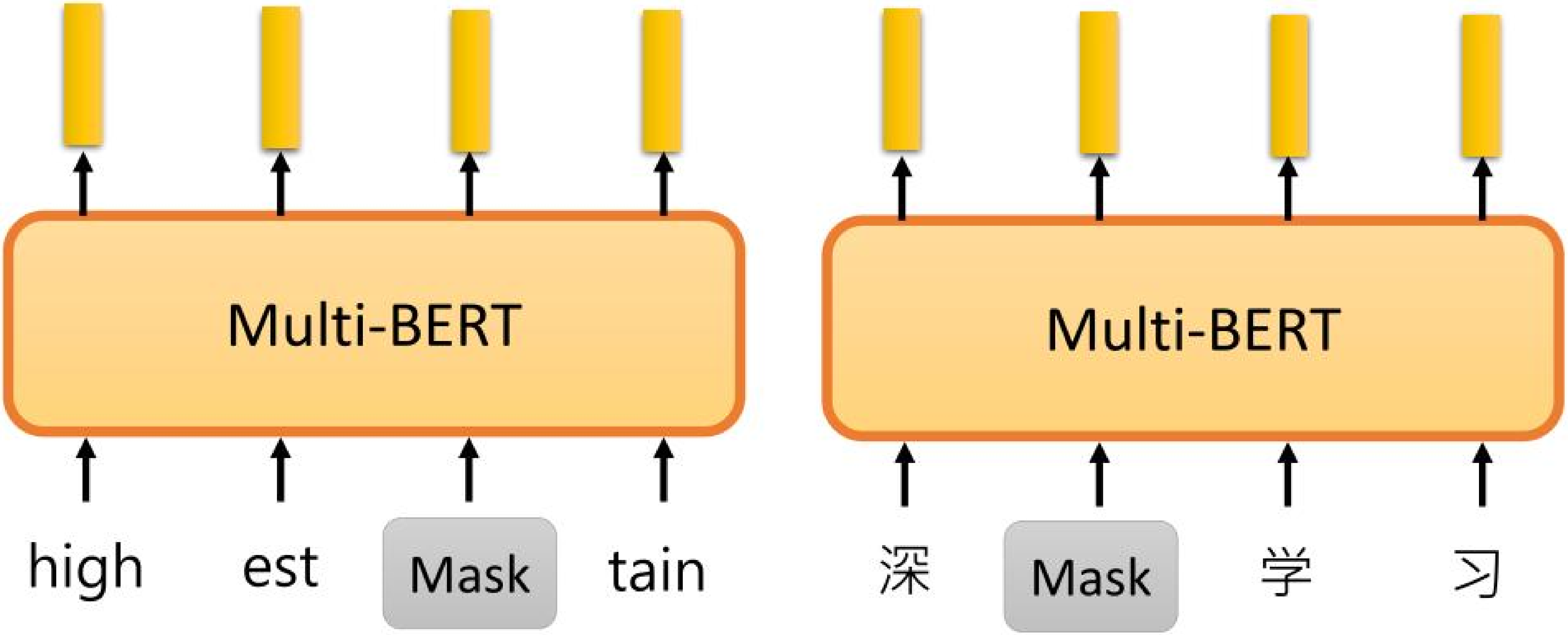
Why does BERT work?

- 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

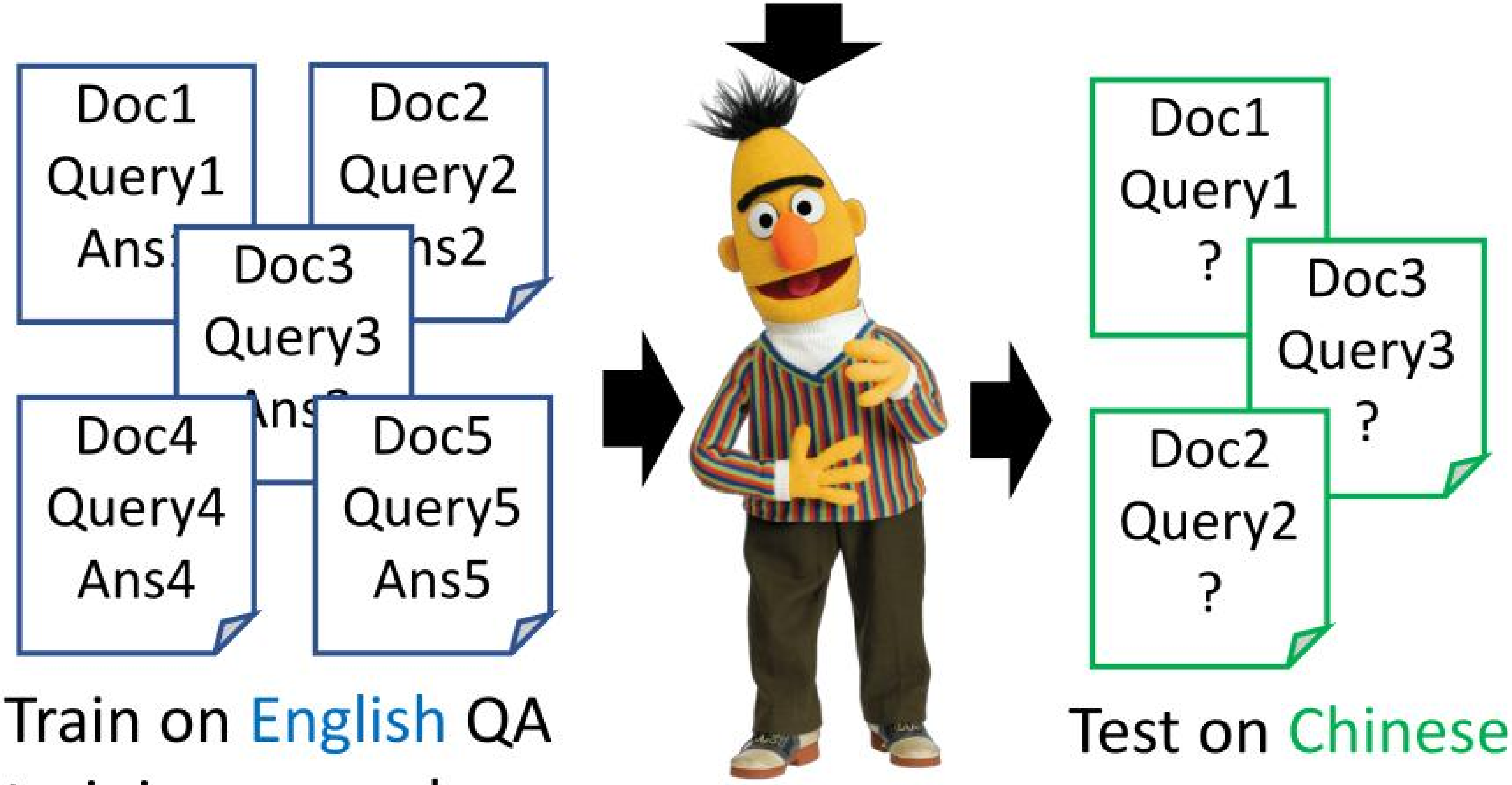
Multi-lingual BERT



Training a BERT model by many different languages.

Zero-shot Reading Comprehension

Training on the sentences of 104 languages



Train on English QA training examples

Multi-BERT

Test on Chinese QA test

Zero-shot Reading Comprehension

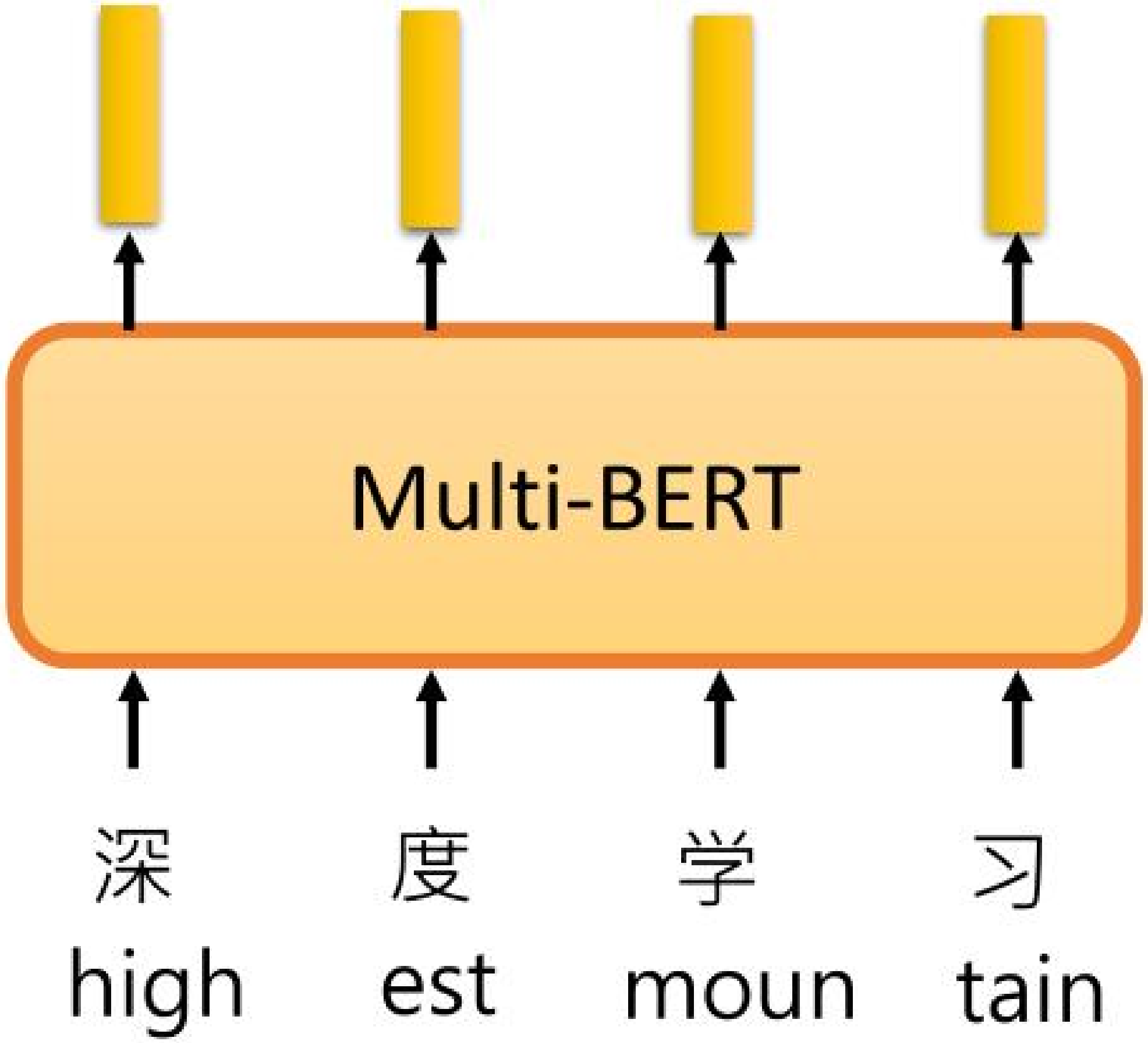
- English: SQuAD, Chinese: DRCD

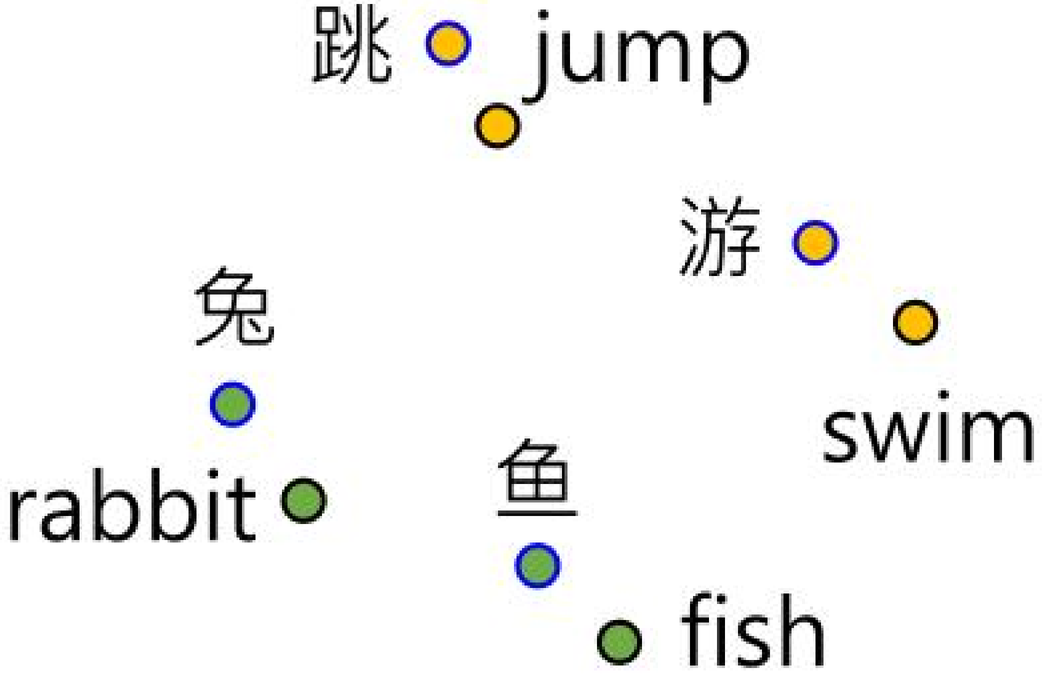
Model	Pre-train	Fine-tune	Test	EM	F1
QANet	none	Chinese	Chinese	66.1	78.1
BERT	Chinese	Chinese		82.0	89.1
	104 languages	Chinese		81.2	88.7
		English		63.3	78.8
		Chinese + English		82.6	90.1

F1 score of Human performance is 93.30%

https://arxiv.org/abs/1909.09587

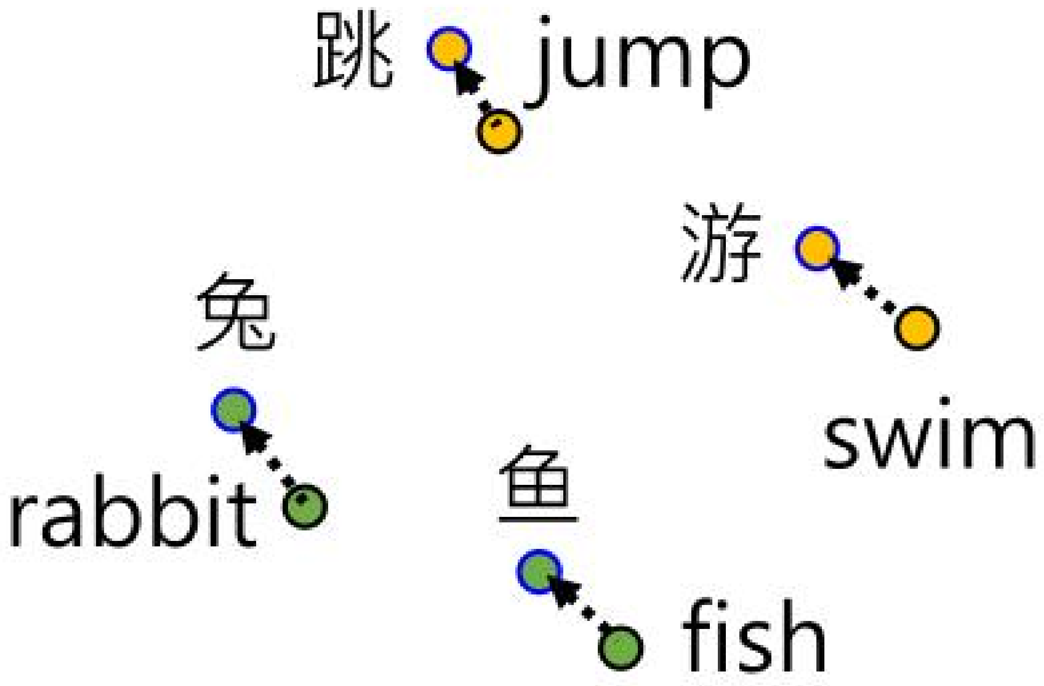
Cross-lingual Alignment?





https://arxiv.org/abs/2010.10041

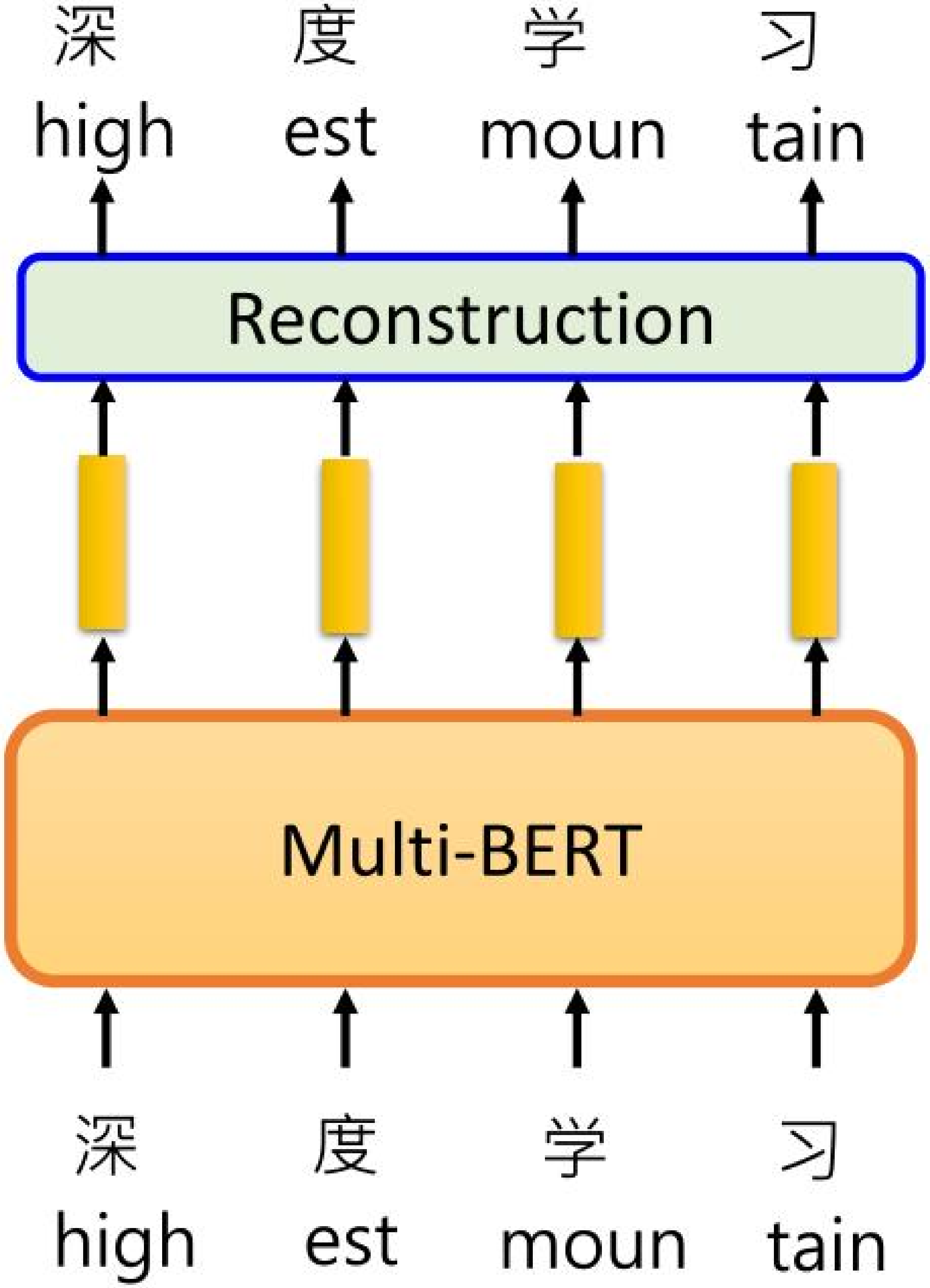
Weird???



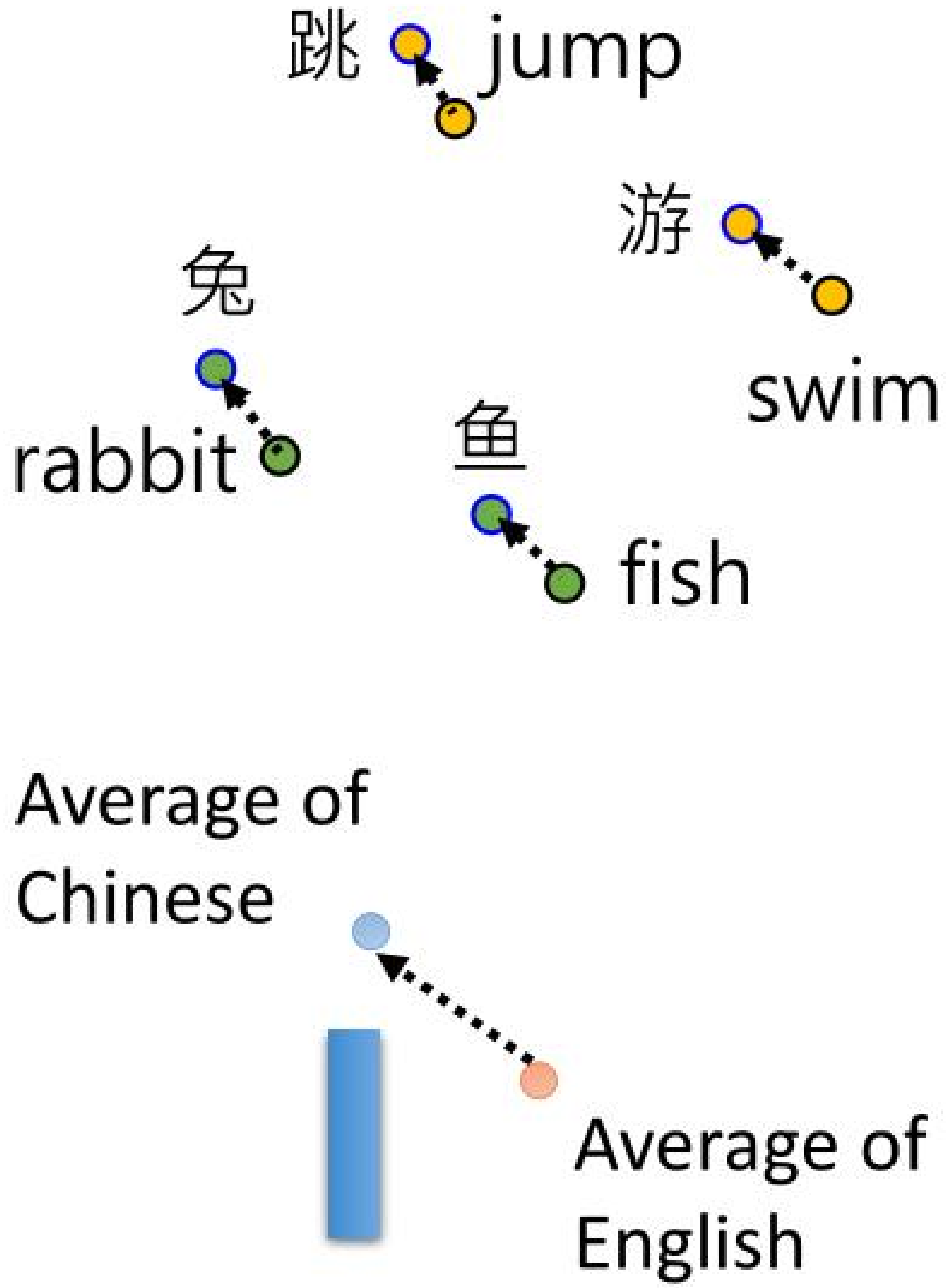
If the embedding is language independent …

How to correctly reconstruct?

There must be language information.



Where is Language?



跳 jump

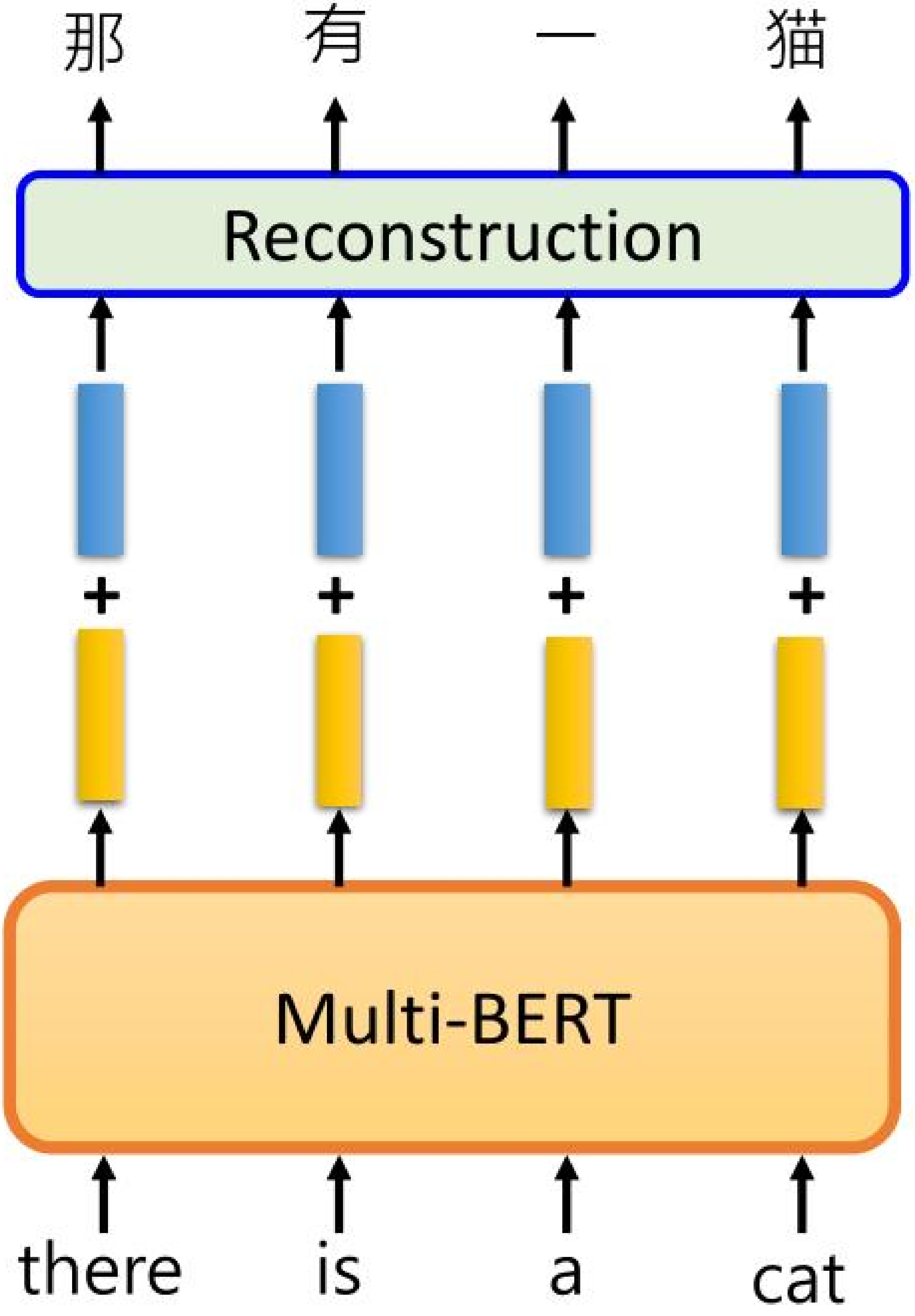
游 swim

兔 rabbit

鱼 fish

Average of Chinese

Average of English



那 有 一 猫

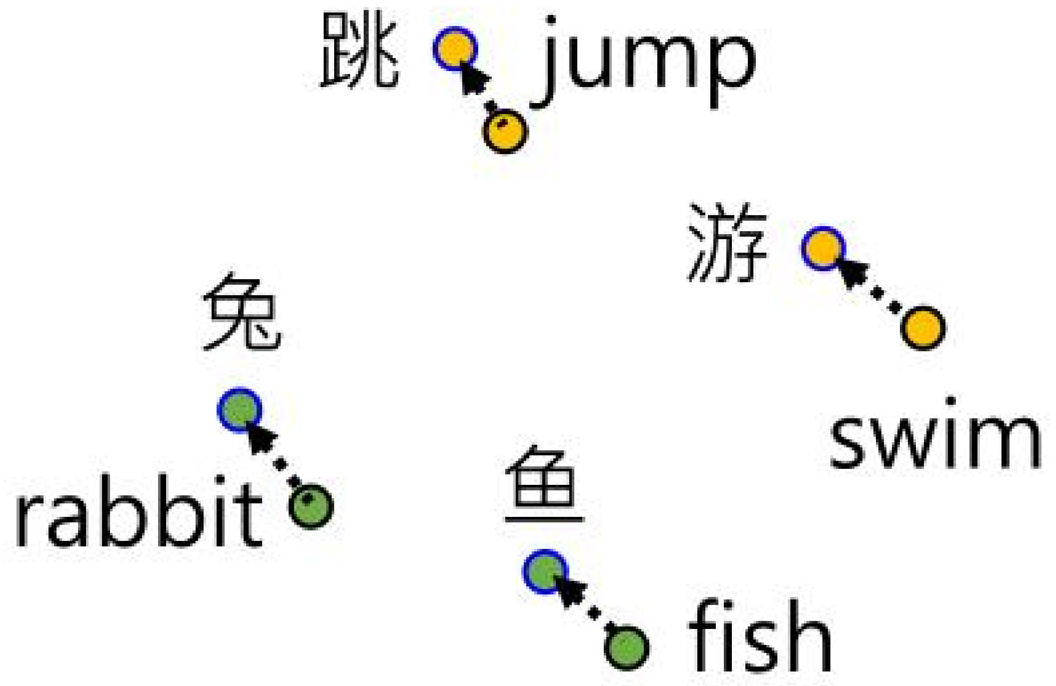
Reconstruction

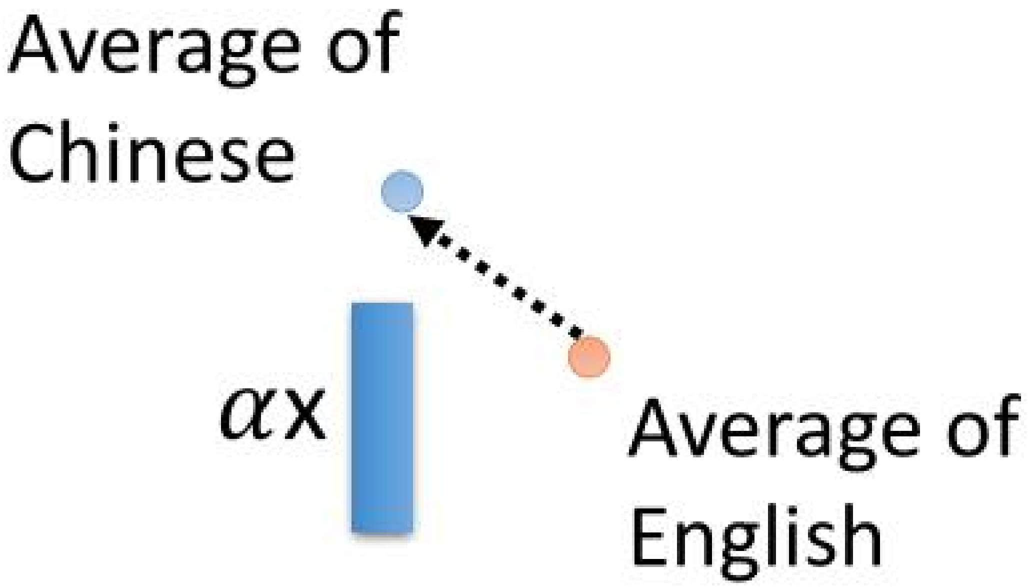
Multi-BERT

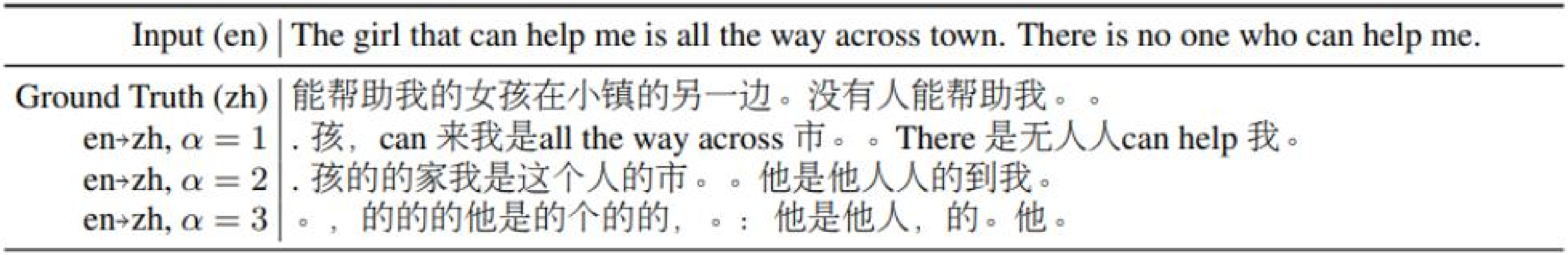
there is a cat

If this is true …

https://arxiv.org/abs/2010.10041







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→zh, $\alpha = 1$	. 孩，can 来我是all the way across 市。。There 是无人人can help 我。
en→zh, $\alpha = 2$	. 孩的的家我是这个人的市。。他是他人人的到我。
en→zh, $\alpha = 3$	。，的的的他是的个的的，。：他是他人，的。他。

Unsupervised token-level translation ☺