Natural Language Processing

第九周 BERT Series

庞彦
yanpang@gzhu.edu.cn

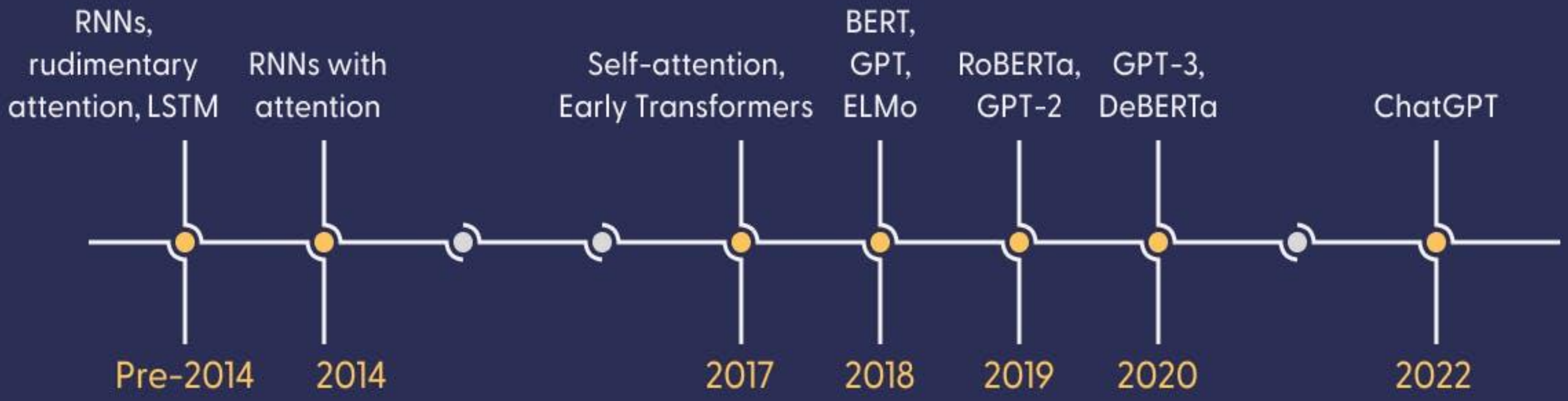
Spring 2023

01

BERT Series

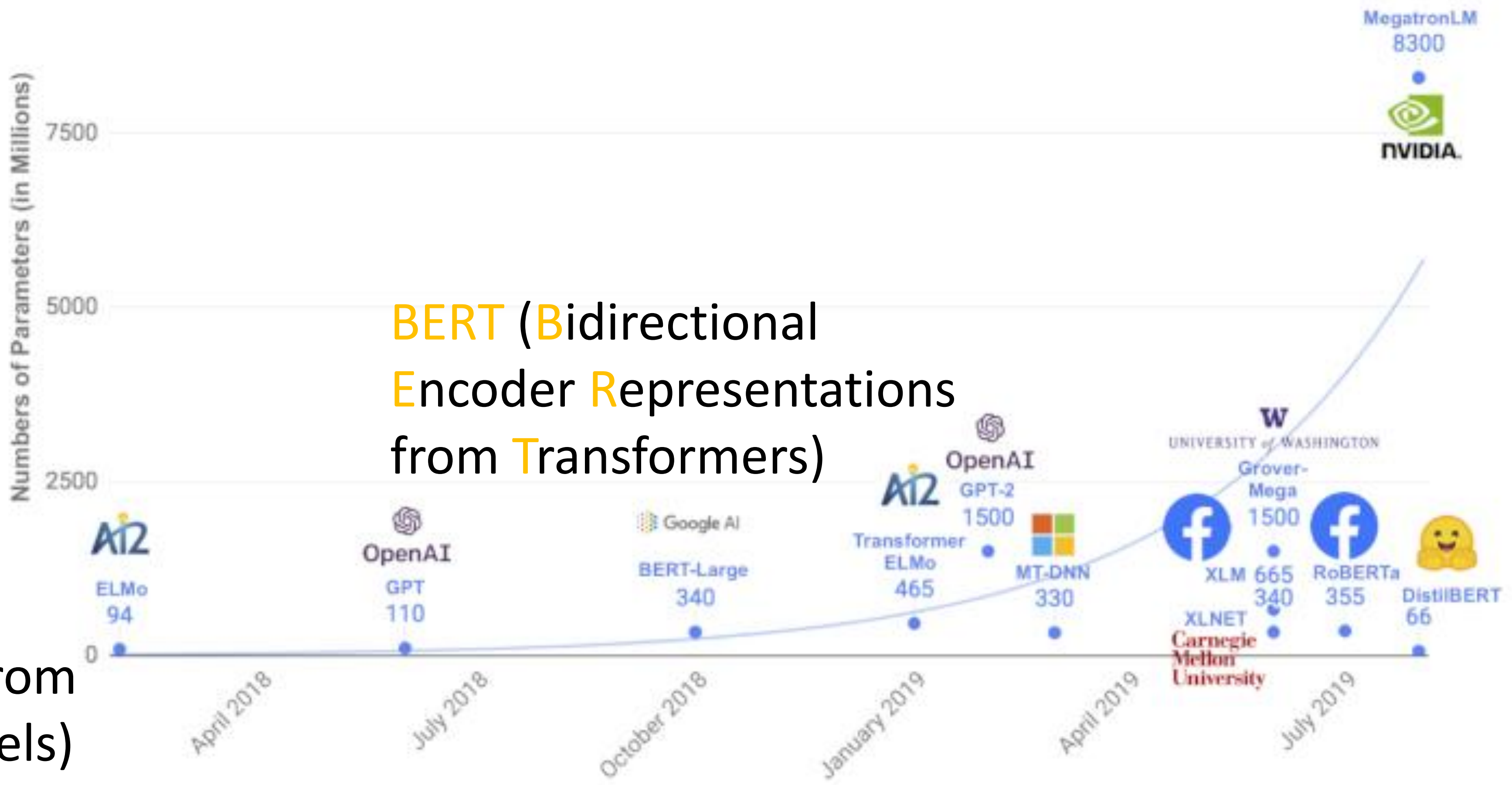
BERT 系列

Timeline



RNNs, rudimentary attention, LSTM
RNNs with attention
Self-attention, Early Transformers
BERT, GPT, ELMo
RoBERTa, GPT-2
GPT-3, DeBERTa
ChatGPT
Pre-2014
2014
2017
2018
2019
2020
2022

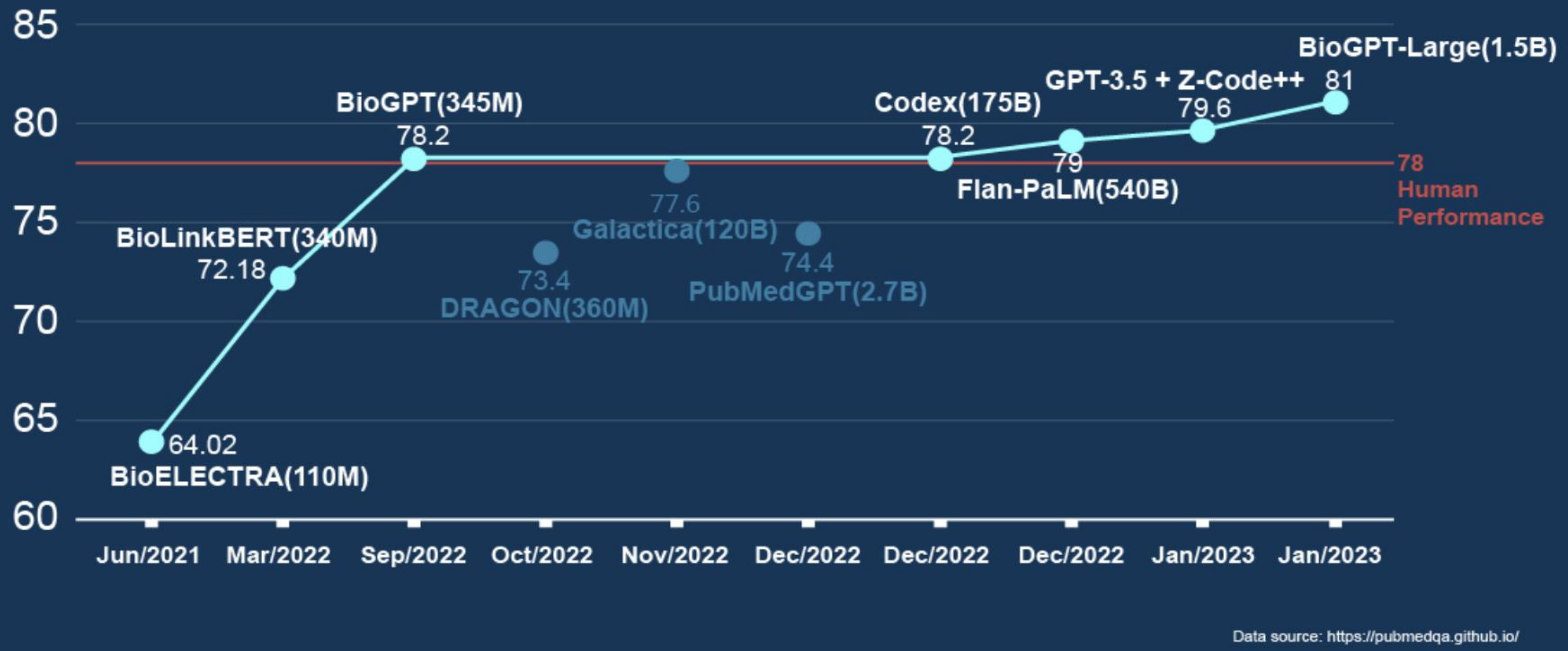
Language Model Size



BERT (Bidirectional Encoder Representations from Transformers)

ELMo (Embeddings from Language Models)

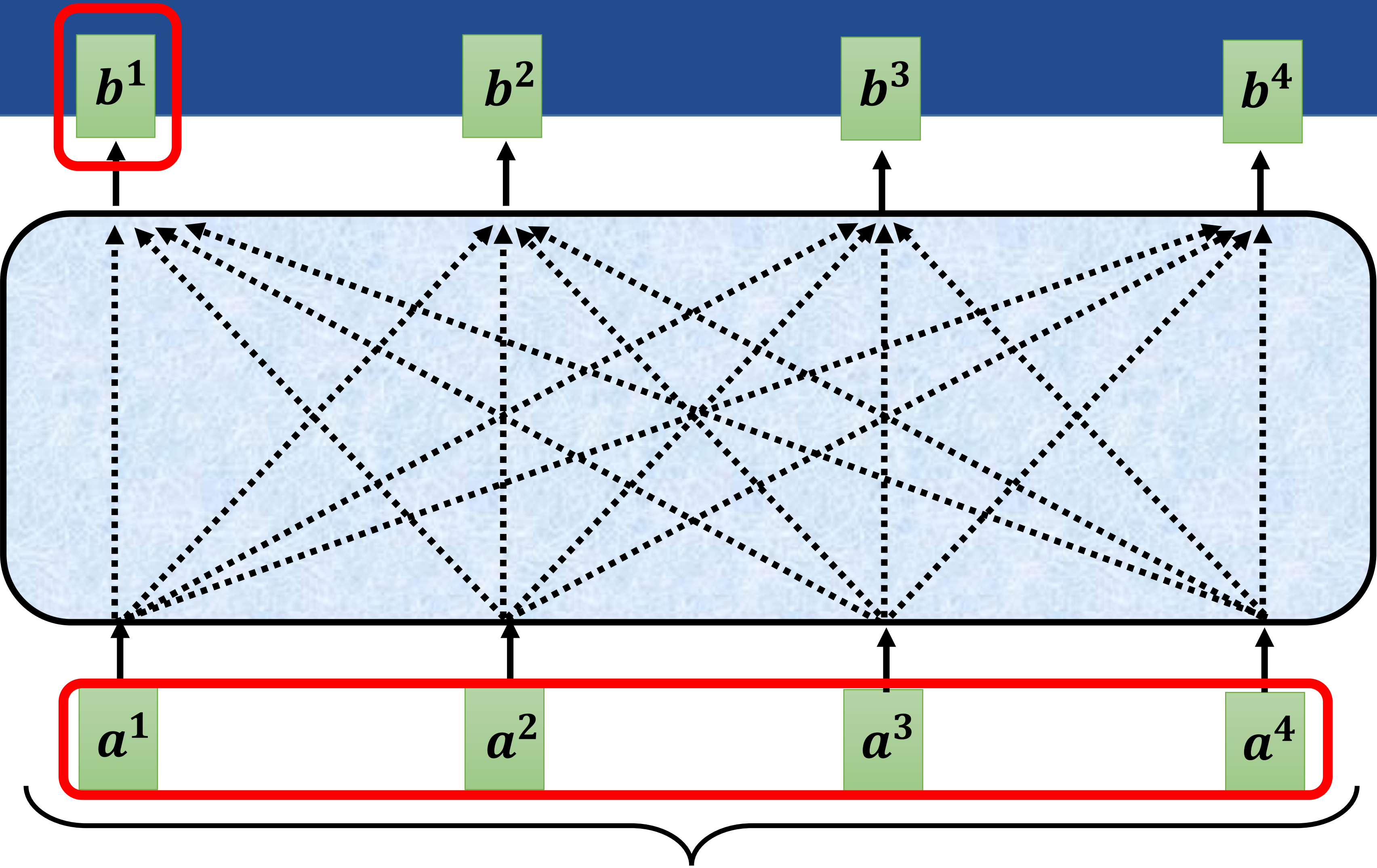
Language Model Size



Overview

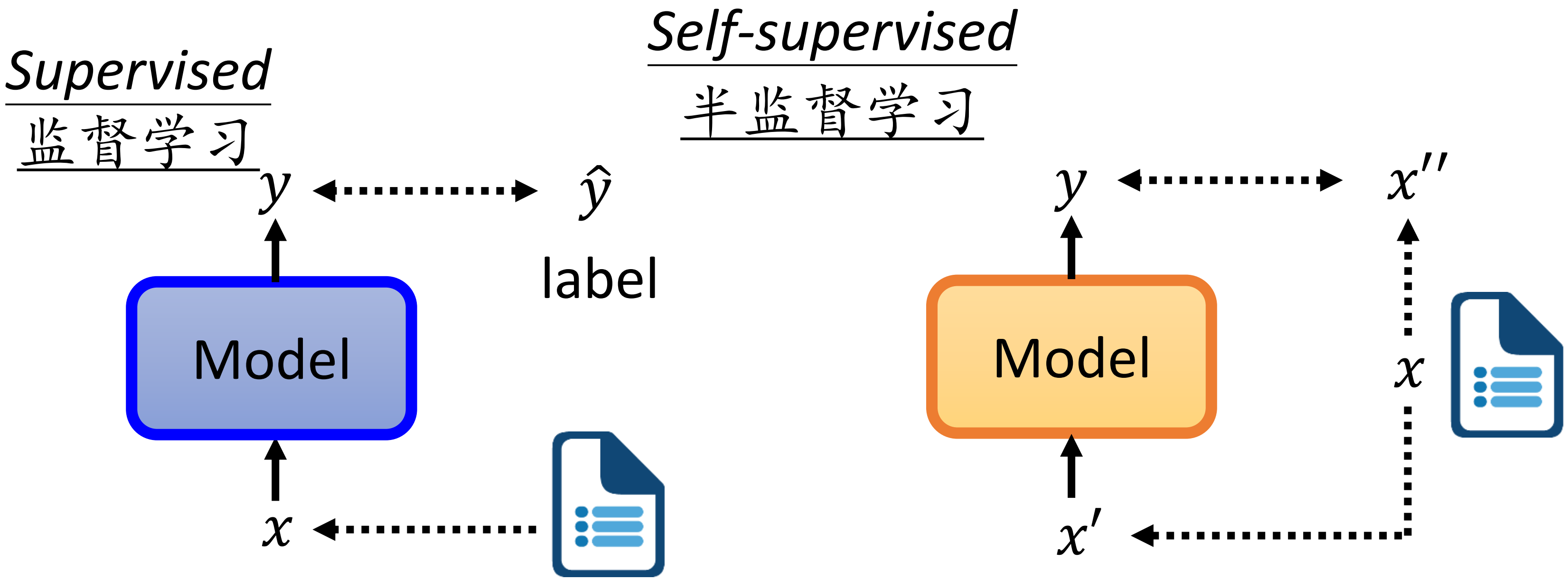
CONTENTS

Review

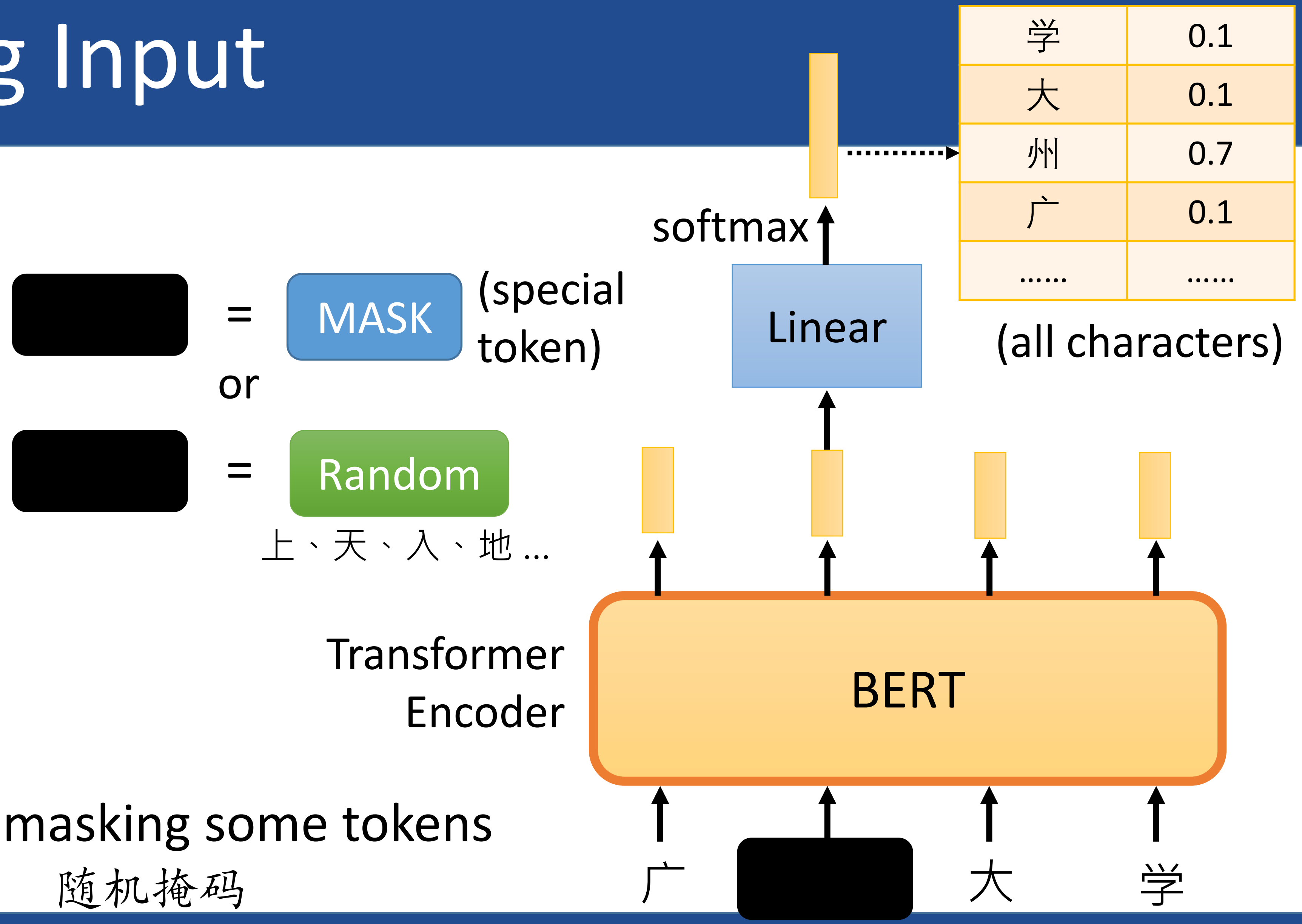


Find the relevant vectors in a sequence找到居中最相关的矢量

Self-supervised Learning



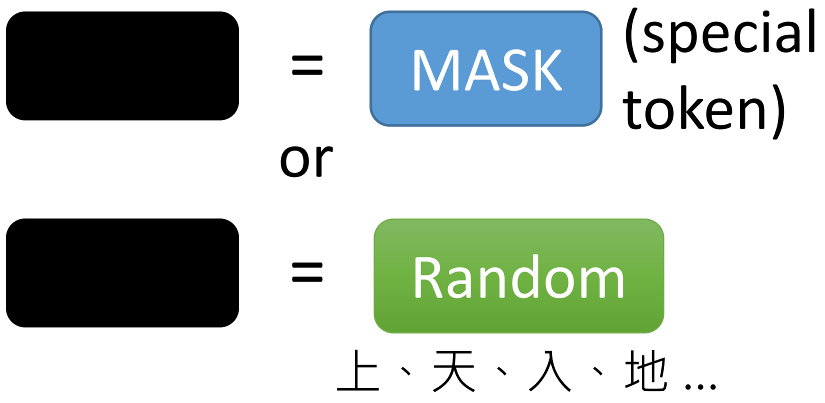
Masking Input

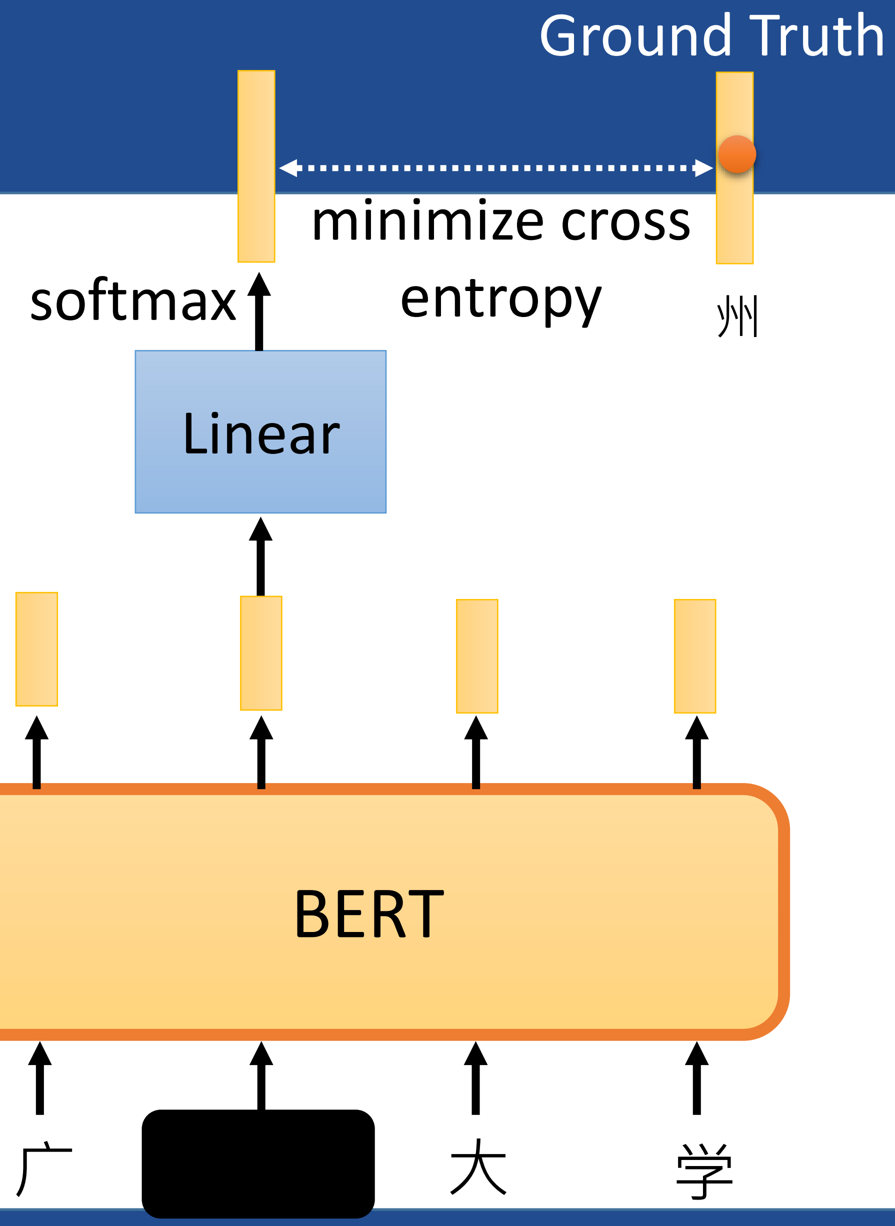


Randomly masking some tokens

随机掩码

Masking Input



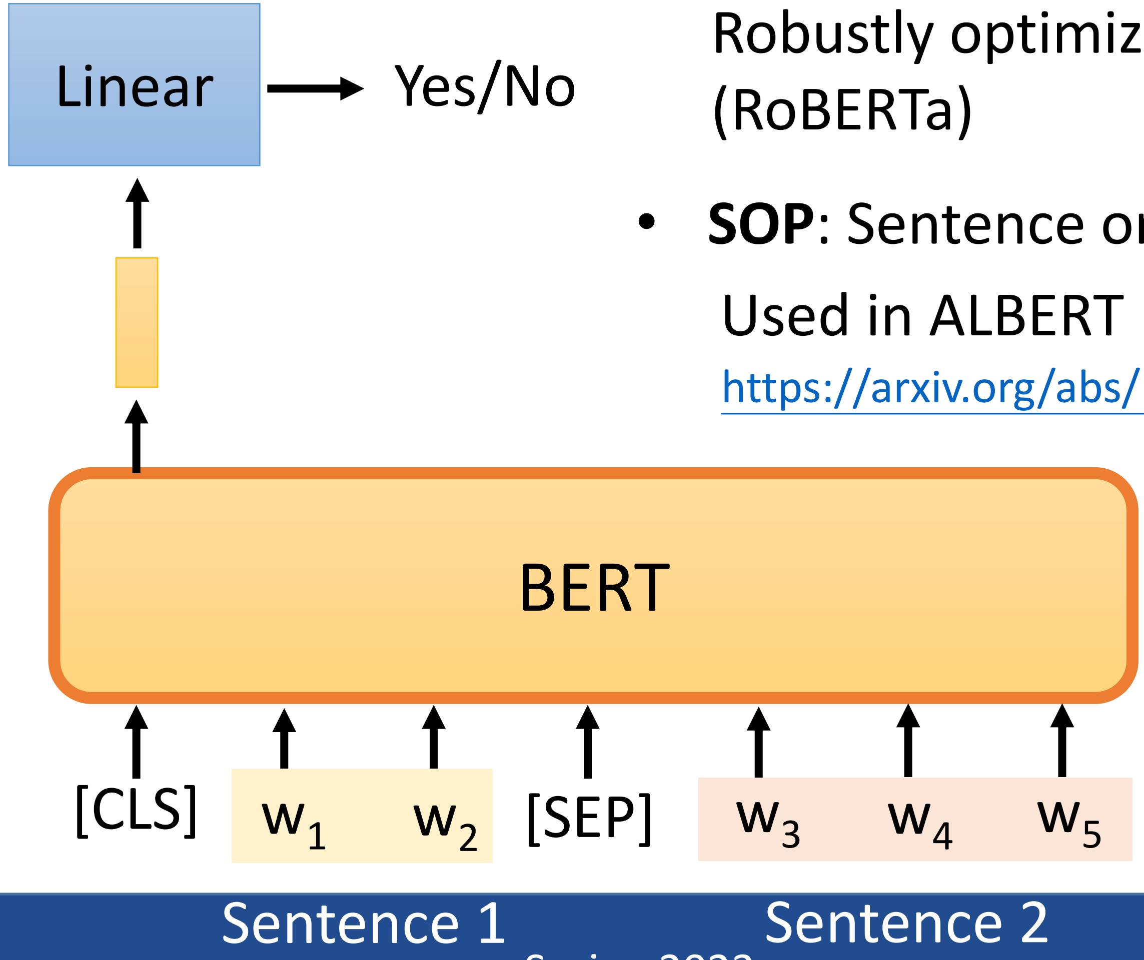


Transformer Encoder

Randomly masking some tokens
随机掩码

https://arxiv.org/abs/1810.04805

Next Sentence Prediction



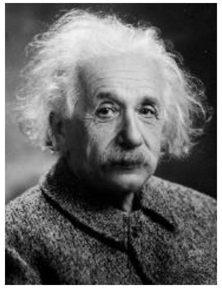
- This approach is not helpful.

  Robustly optimized BERT approach (RoBERTa)

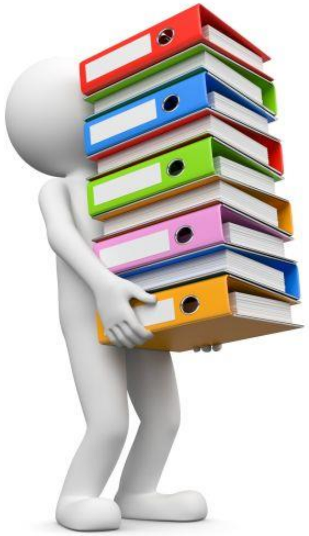
- **SOP**: Sentence order prediction

  Used in ALBERT

  https://arxiv.org/abs/1909.11942



Next Sentence Prediction



Self-supervised Learning

Pre-train

BERT

- Masked token prediction
- Next sentence prediction

微调 **Fine-tune**

Downstream Tasks
下游任务

Model for Task 1

Model for Task 2

Model for Task 3

GLUE

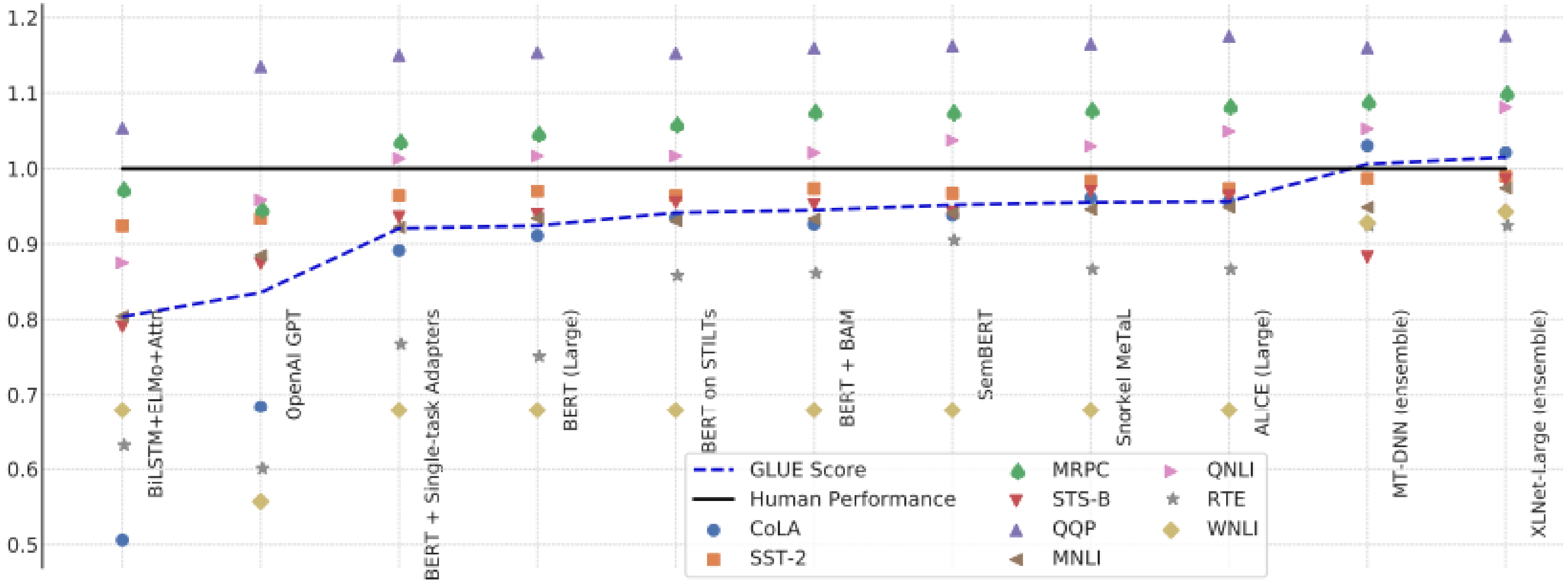
- Corpus of Linguistic Acceptability (CoLA)
- Stanford Sentiment Treebank (SST-2)
- Microsoft Research Paraphrase Corpus (MRPC)
- Quora Question Pairs (QQP)
- Semantic Textual Similarity Benchmark (STS-B)
- Multi-Genre Natural Language Inference (MNLI)
- Question-answering NLI (QNLI)
- Recognizing Textual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

General Language Understanding Evaluation (GLUE)

Chinese Version: https://www.cluebenchmarks.com

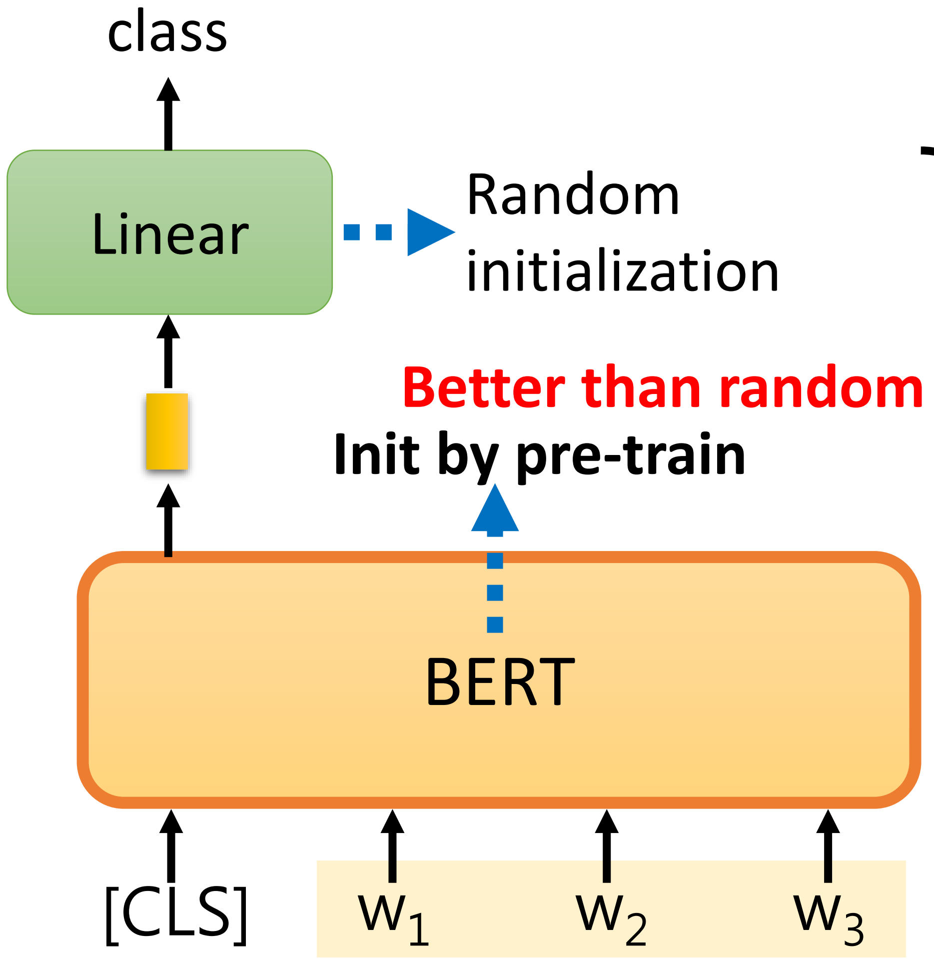
https://gluebenchmark.com

BERT and its Family



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

How to use BERT – Case 1



Input: sequence 句子为输入
output: class 类别为输出

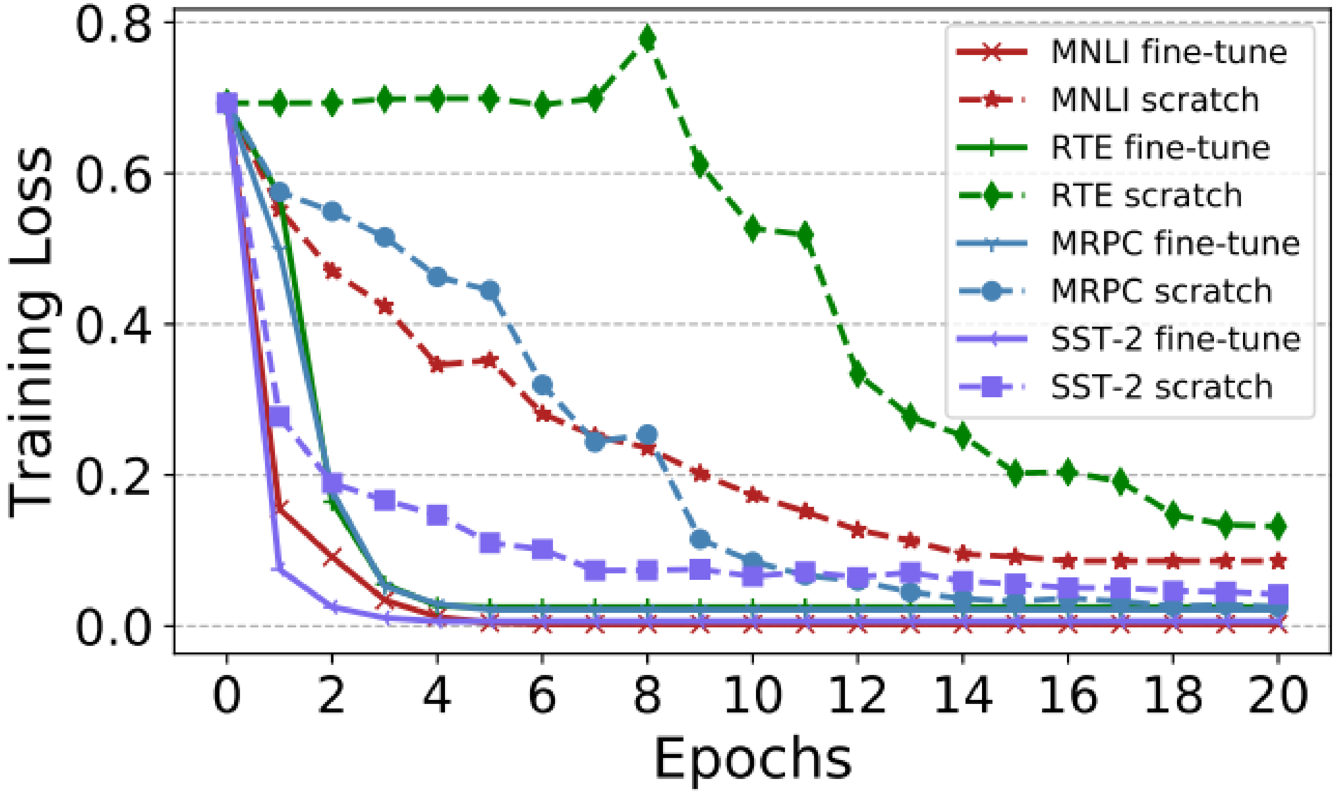
Example:
Sentiment analysis 情感分析

this is good
positive

This is the model to be learned.

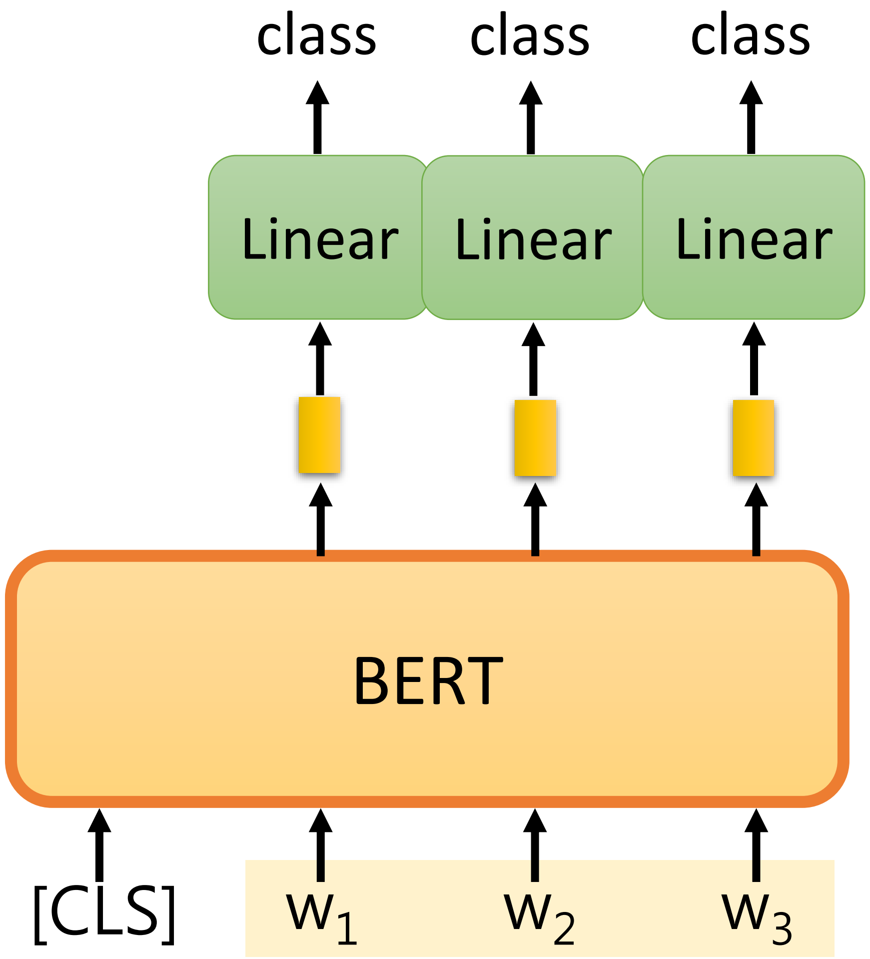
Pre-train v.s. Random Initialization

(fine-tune) (scratch)

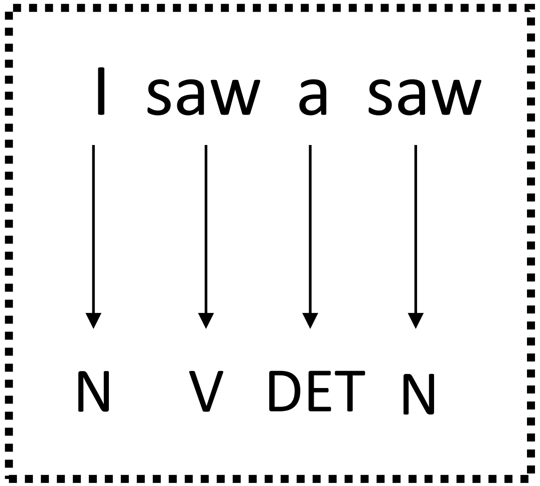


Source of image: https://arxiv.org/abs/1908.05620

How to use BERT – Case 2



Input: sequence
output: same as input
输入输出长度一样
Example:
POS tagging



How to use BERT – Case 3

Input: two sequences 输入两句话
Output: a class 输出一个类

Example:
Natural Language Inferencee (NLI)

contradiction
entailment
neutral

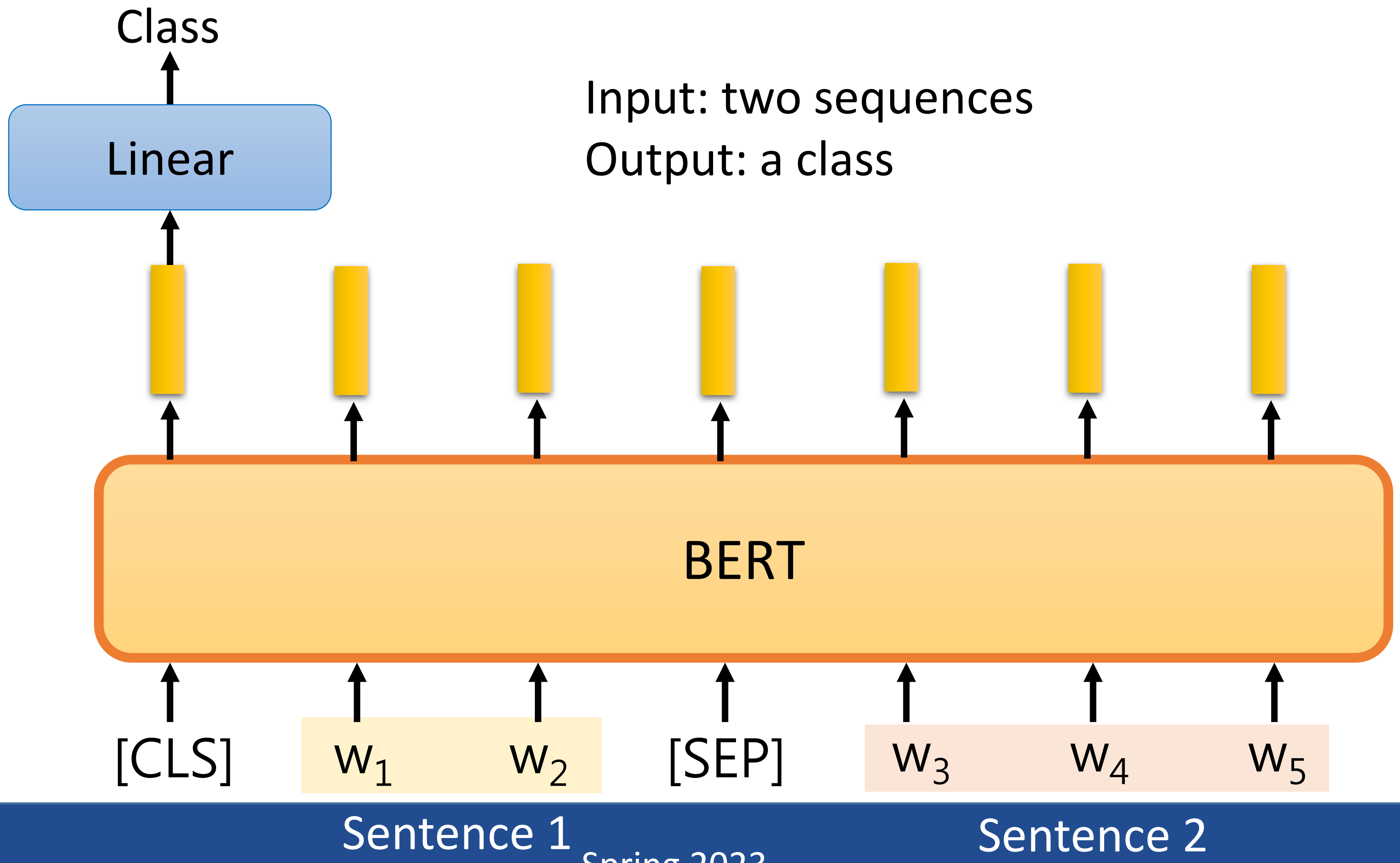
Model

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

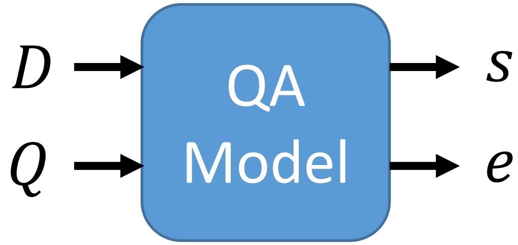
Output: a class

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".

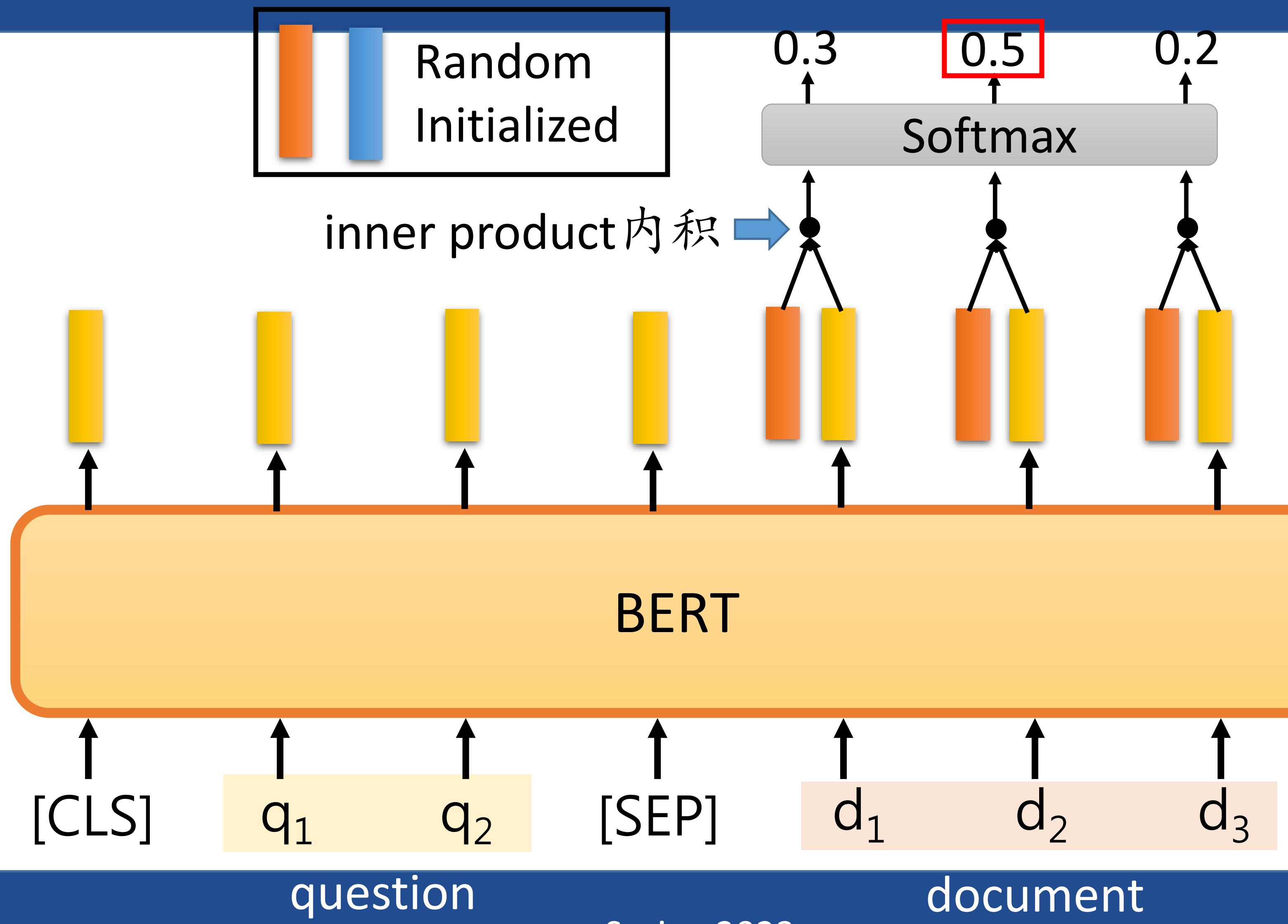
What causes precipitation to fall?
gravity $s = 17, e = 17$

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

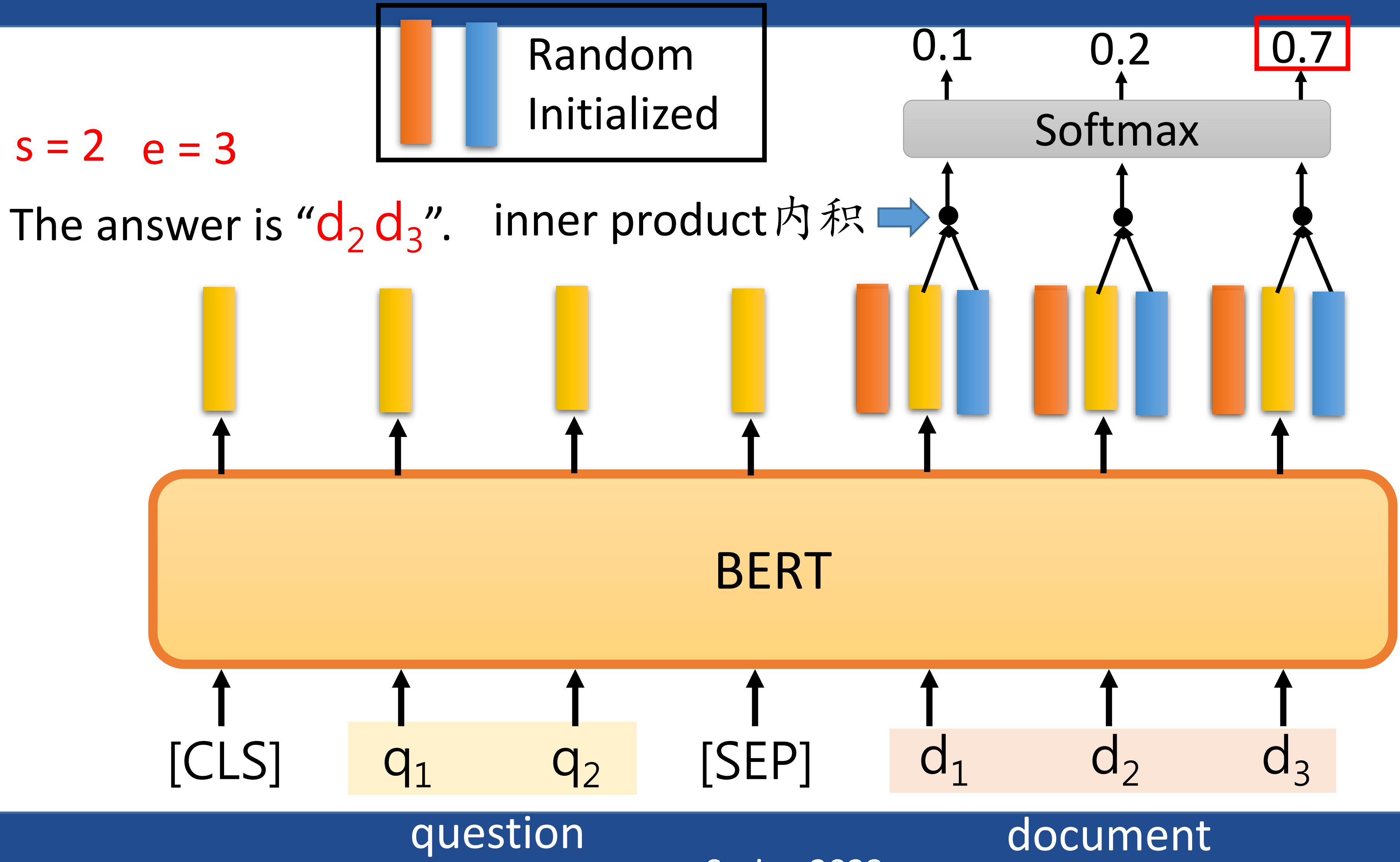
Where do water droplets collide with ice crystals to form precipitation?
within a cloud $s = 77, e = 79$

How to use BERT – Case 4

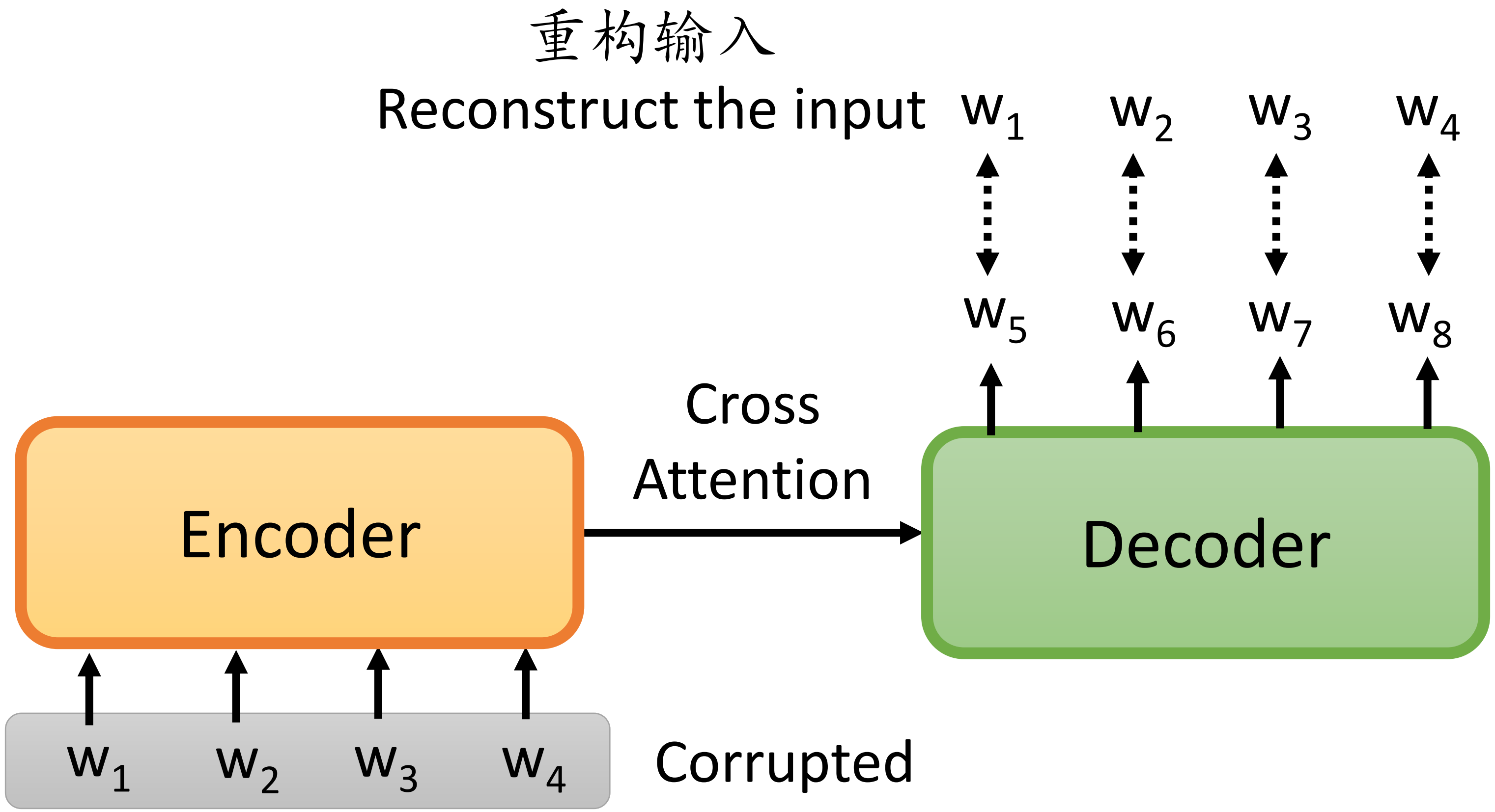
s = 2



How to use BERT – Case 4

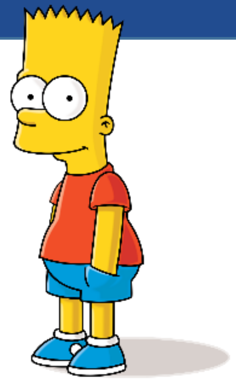


Pre-training a seq2seq model



MASS / BART

BART



A B [SEP] C D E

A B [SEP] C ▢ E

MASS

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

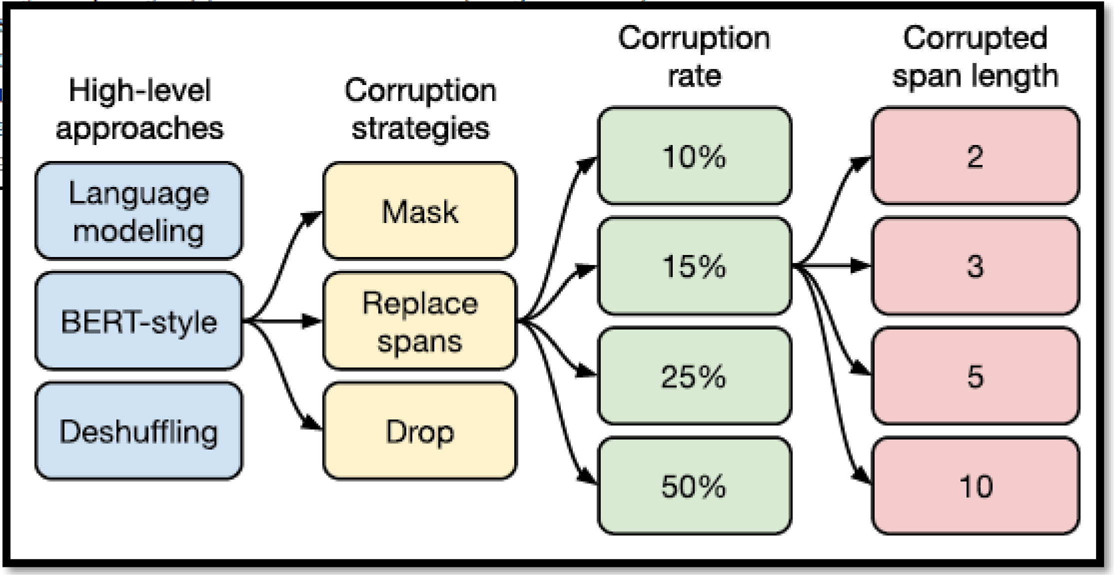
Text Infilling

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

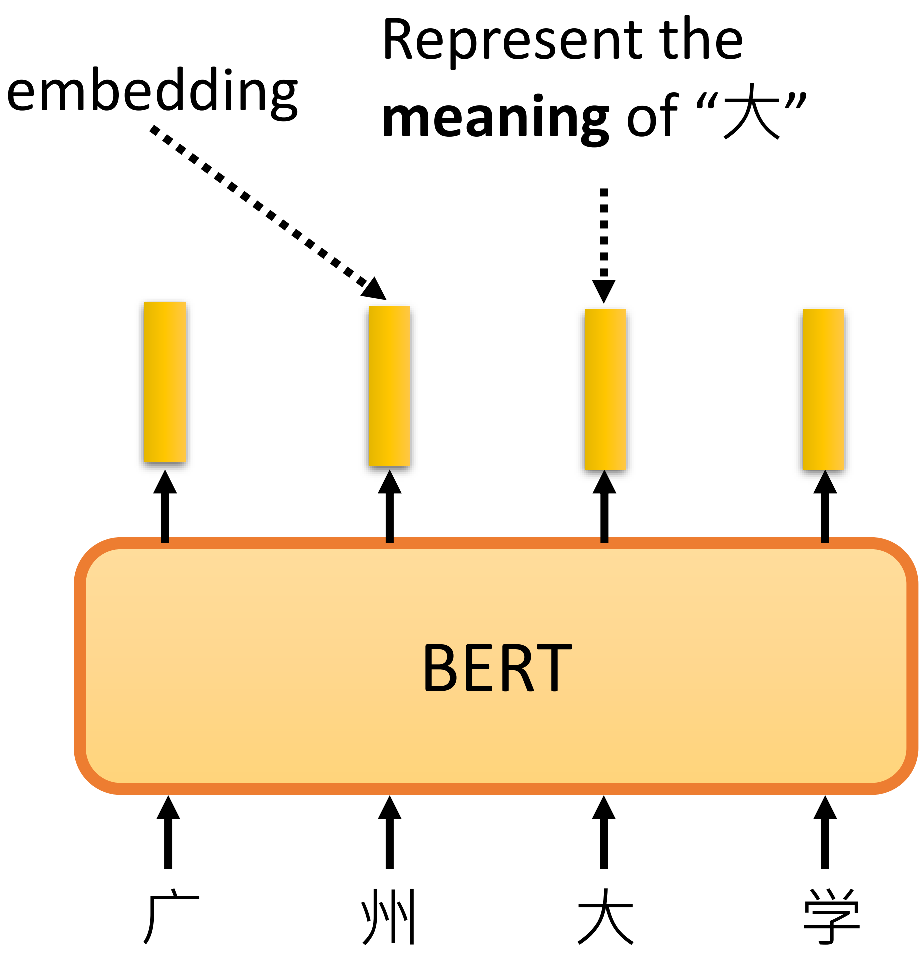
T5 – Comparison

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

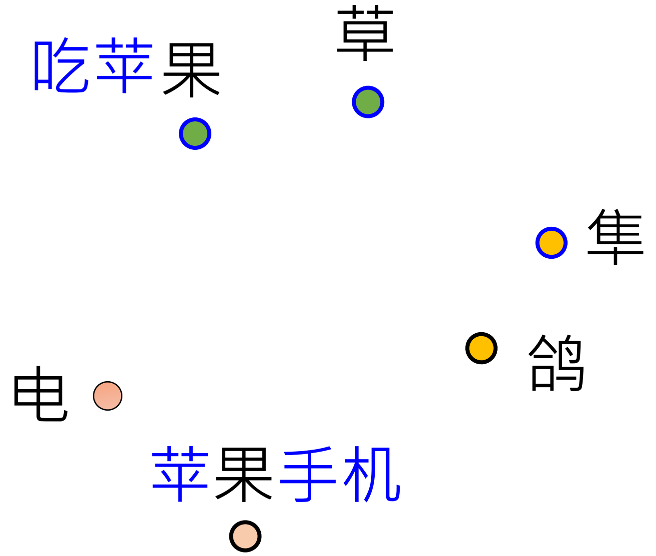
Objective	Inputs	Targets
Prefix language modeling	Thank you for inviting	me to your party last week .
BERT-style	Thank you <M> <M> me to your party apple week .	*(original text)*
Deshuffling	party me for your to . las	
I.i.d. noise, mask tokens	Thank you <M> <M> me to	
I.i.d. noise, replace spans	Thank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



Why does BERT work?

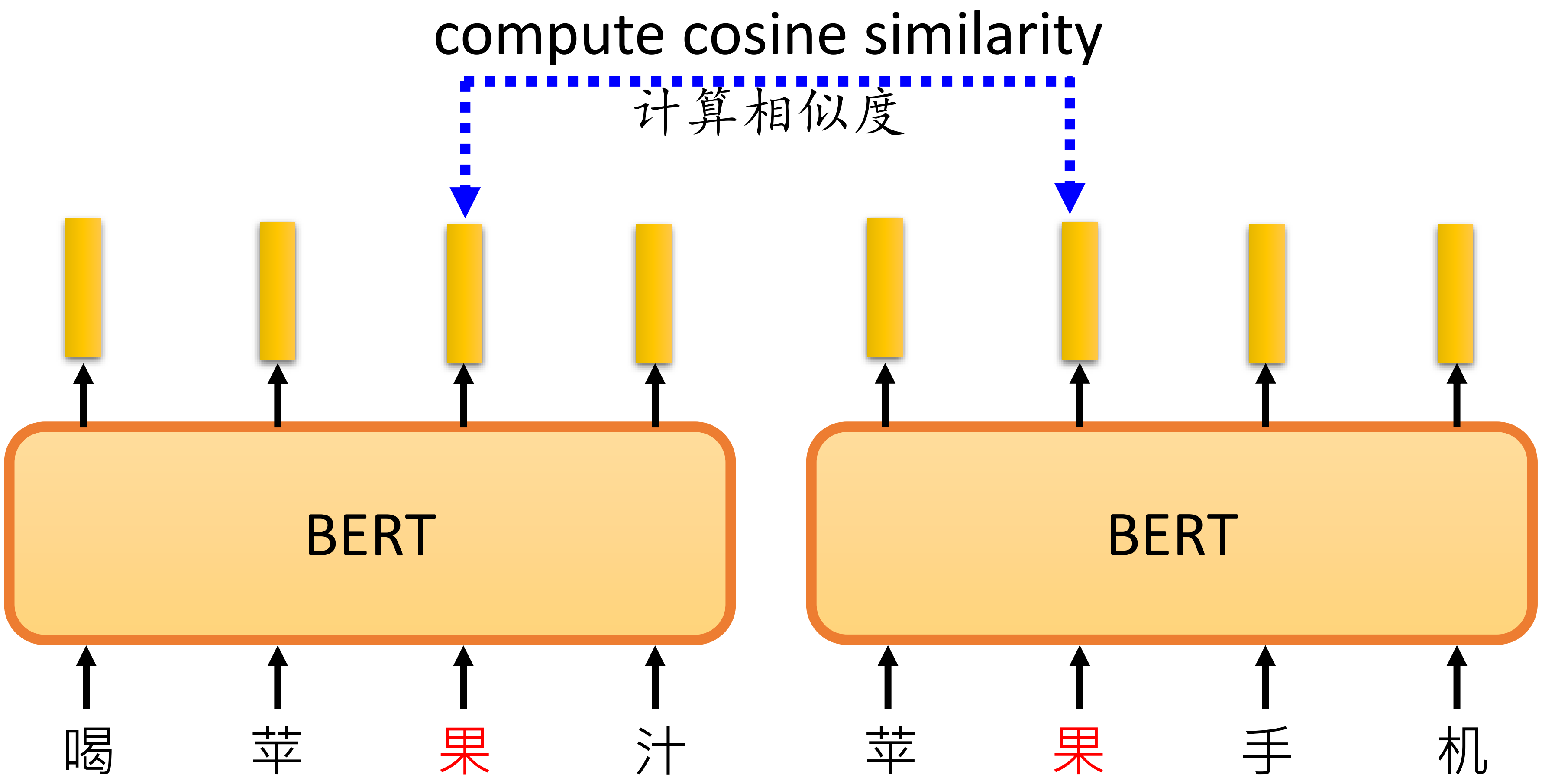


The tokens with similar meaning have similar embedding.
近义词含有类似的表征。



Context is considered.

Why does BERT work?



Similarity

内积相似度：

$$sim(q,k) = q^T k$$

余弦相似度：

$$sim(q,k) = \frac{q^T k}{\|q\| \|k\|}$$

拼接相似度：

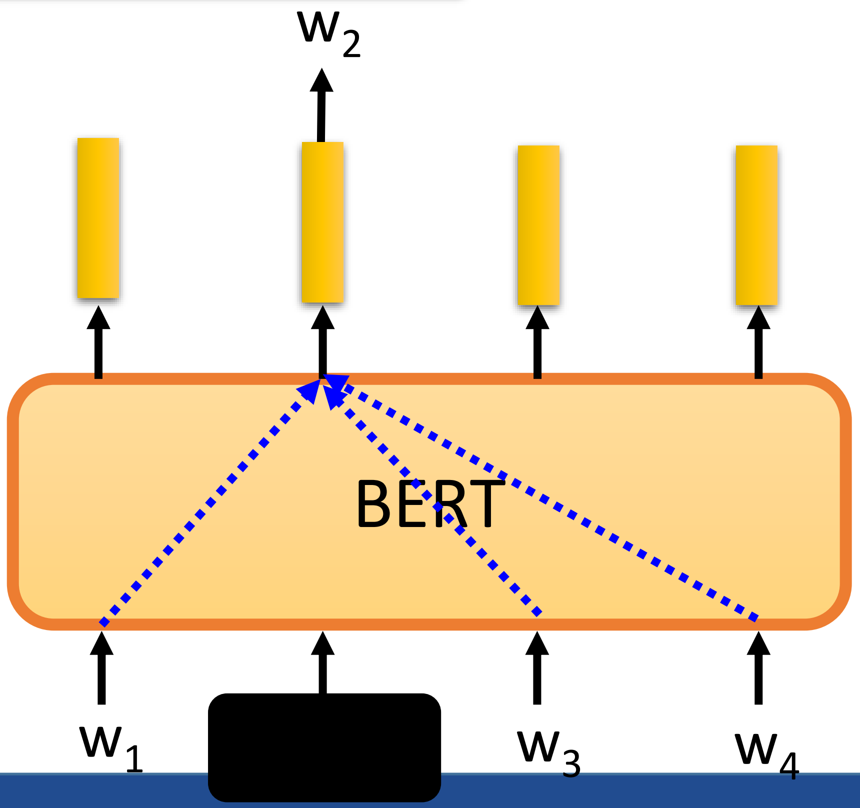
$$sim(q,k) = \omega^T k[q;k] = \omega_1^T q + \omega_2^T k$$

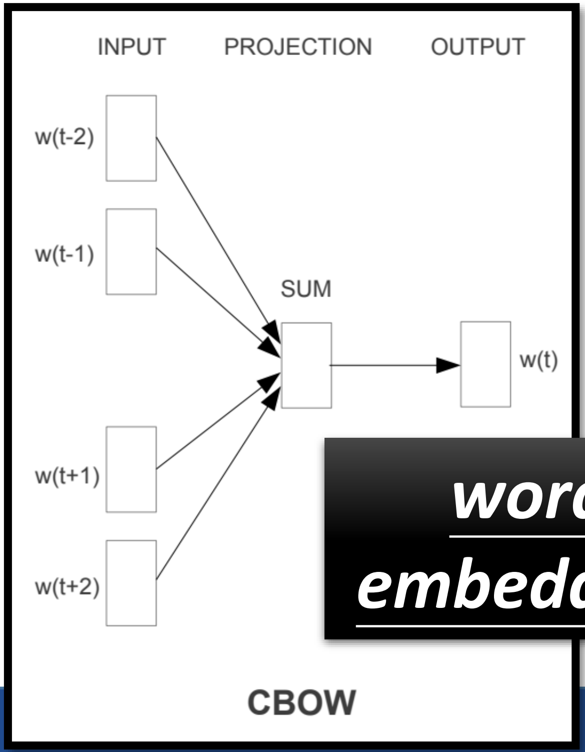
Why does BERT work?

Contextualized word embedding

You shall know a word by the company it keeps

John Rupert Firth





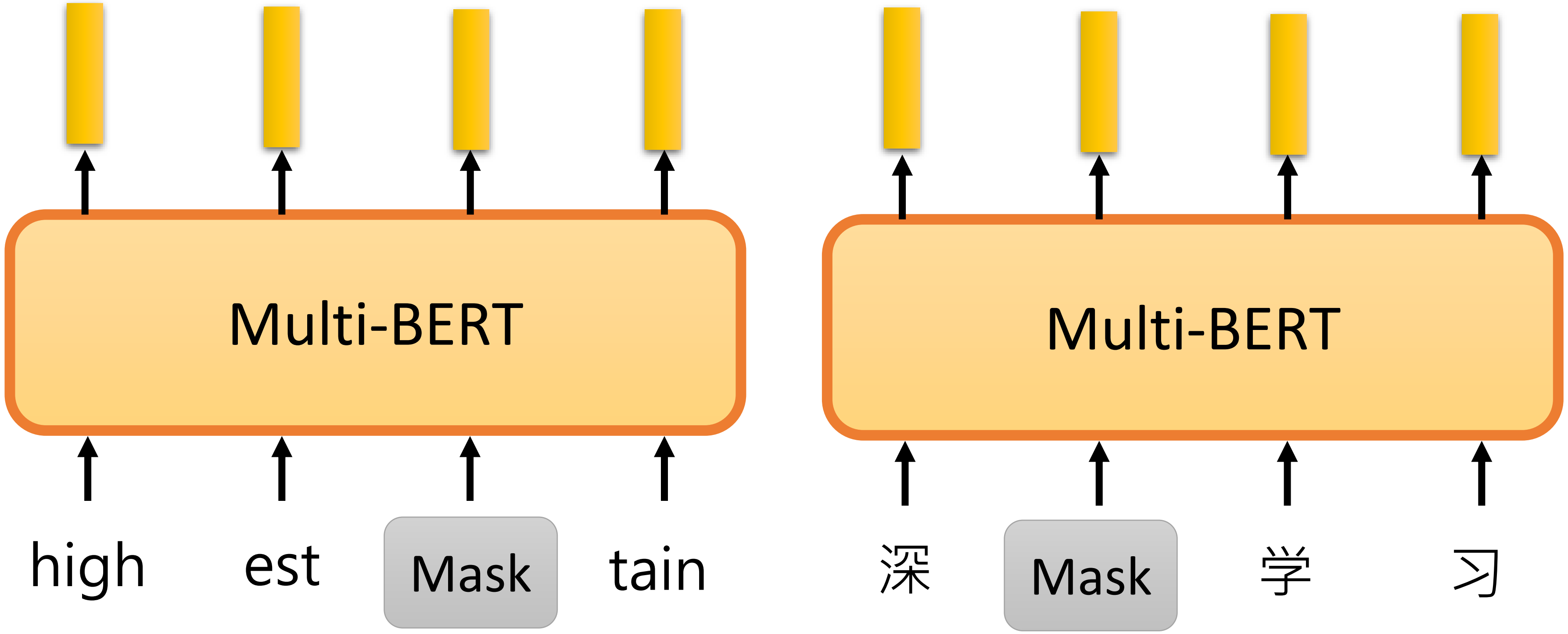
word embedding

DNA Sequence


```
                0000000000000000000111111111111111111111
                88888888899999999990000000000111111111122
                12345678901234567890123456789012345678901
Taxon 1         GCCTAGCCAAAGCTCTTCCAAGGTGACTCTCAGTTCAAGCT
Taxon 2         GCCTAGCCAAAGCTCTTCCAAGCTGACTCTCA------GCT
Taxon 3         GCCTAGCCTAAGCTCAACCAAGGTGTCTCTCAGTTCAAGCT
Taxon 4         GCCTAGCCTAAGCTCTTCCAAGGTGTCTCTCAGTTCAAGCT
Taxon 5         GCCTAGCCAAAGCTCTTCCAAGCTGACTCTCA------GCT
Taxon 6         CCCTAGCCAAAGCTCTTCCAAGCTGACTCTCAGTTCAAGCT
Taxon 7         CCCTAGCCAAAGCTCTTCCAAGCTGACTCTCAGTTCAAGCT
Taxon 8         GCCTAGCCTAAGCTCTTCCAAGCTGACTCTCAGTTCAAGCT
```

类别　　　　　　　　DNA 序列

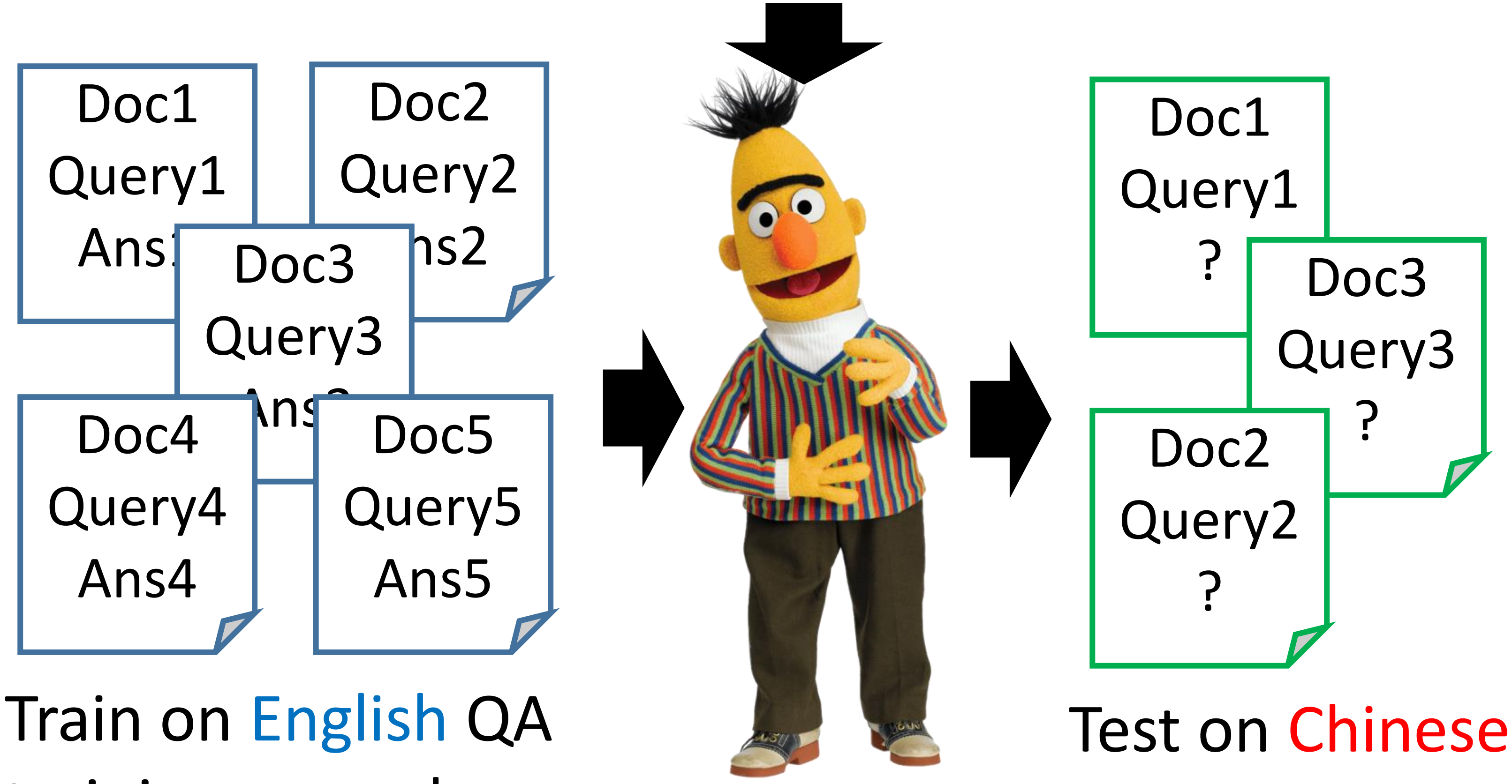
Multi-lingual BERT



Training a BERT model by many different languages.
各种语言都可以利用BERT来训练

Zero-shot Reading Comprehension

Training on the sentences of 104 languages 104种语言训练



Train on English QA training examples

Multi-BERT

Test on Chinese QA test

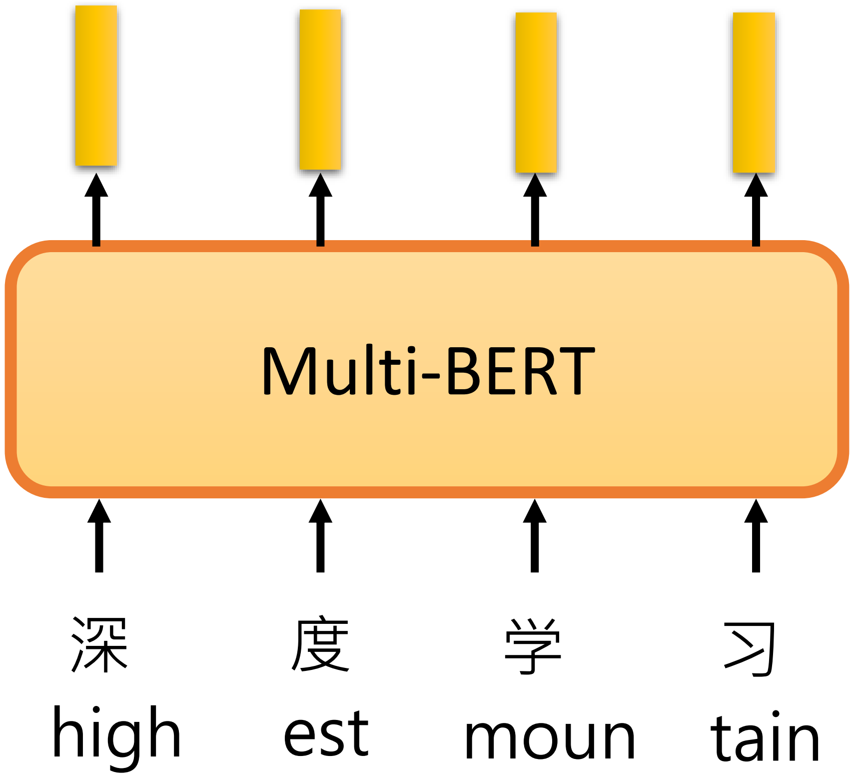
Zero-shot Reading Comprehension

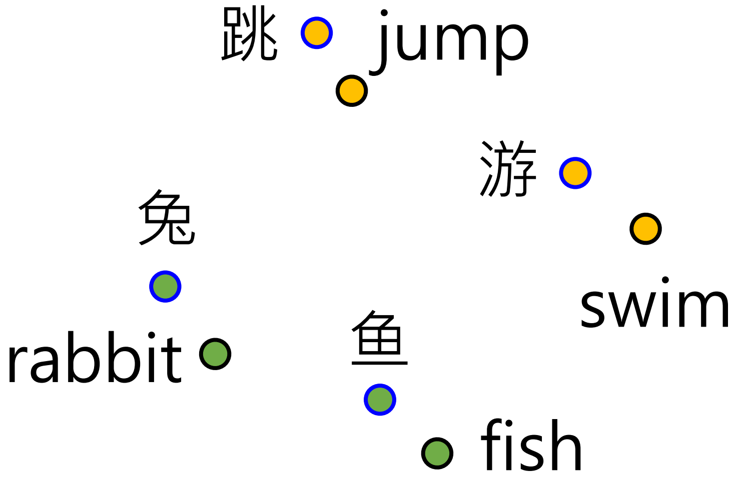
English：SQuAD; Chinese：DRCD

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%

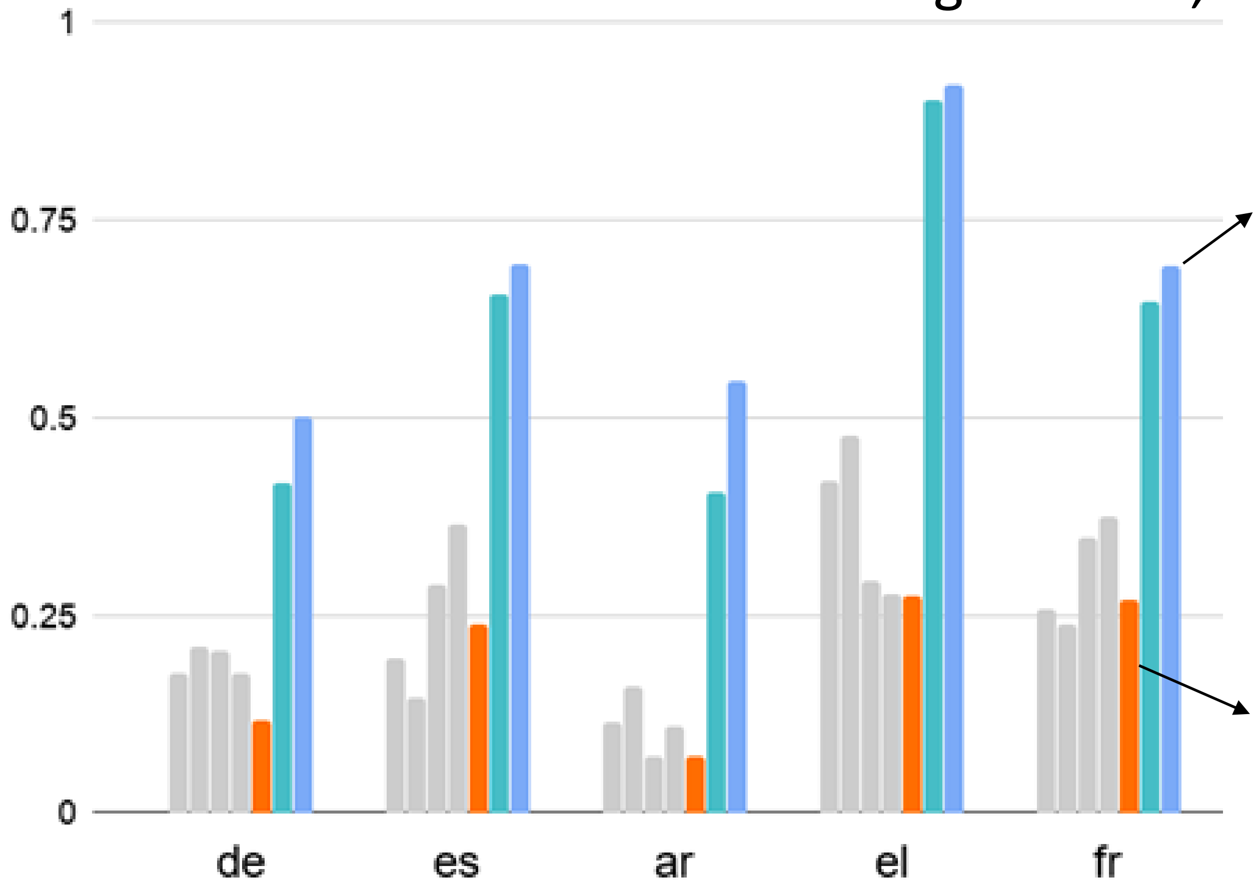
Cross-lingual Alignment





Mean Reciprocal Rank (MRR)

Higher MRR, better alignment

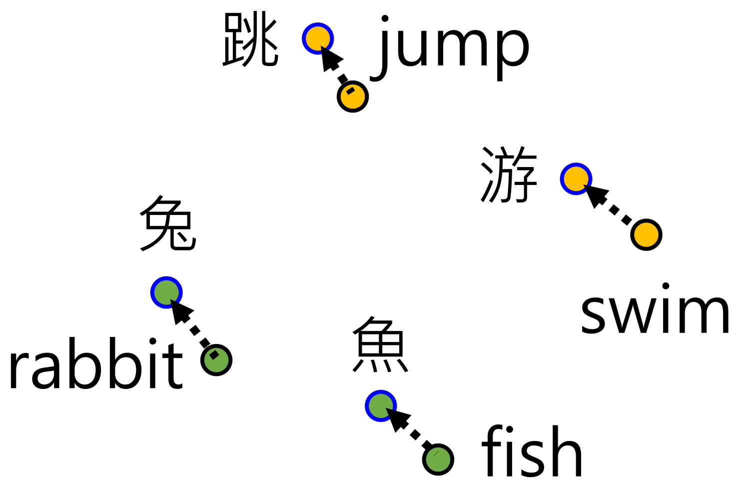


Google's Multi-BERT

Our Multi-BERT

200k sentences for each lang

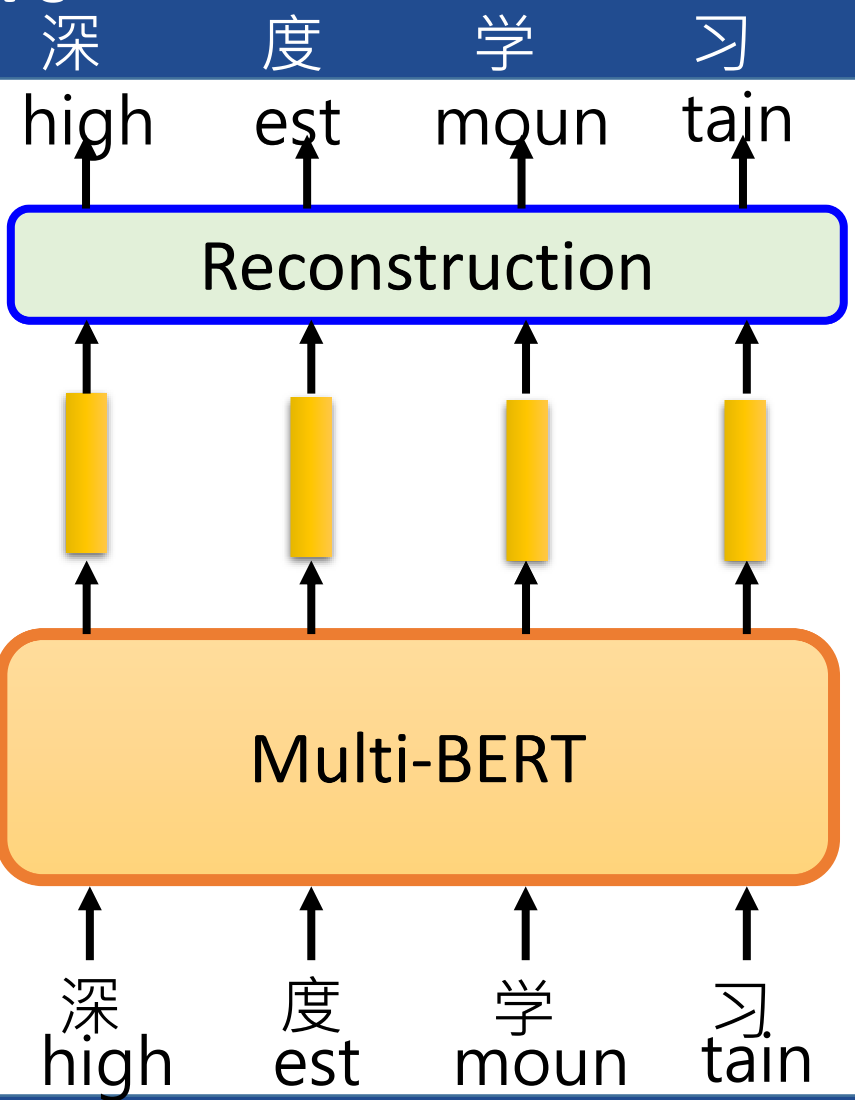
Cross-lingual Alignment



If the embedding is language independent …

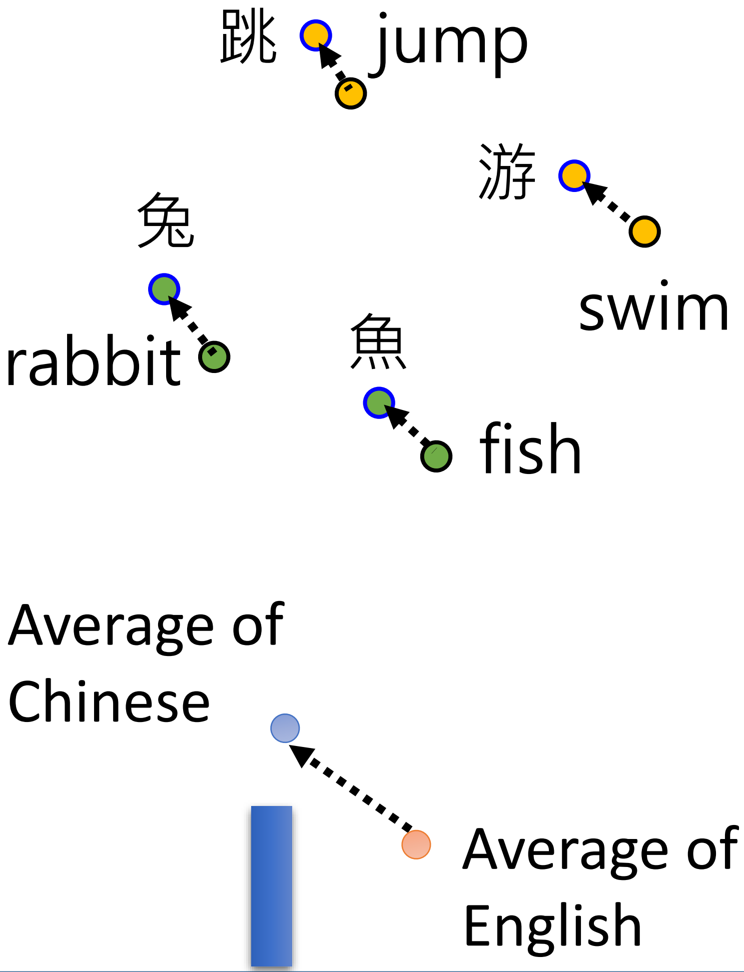
How to correctly reconstruct? 如何正确重构？

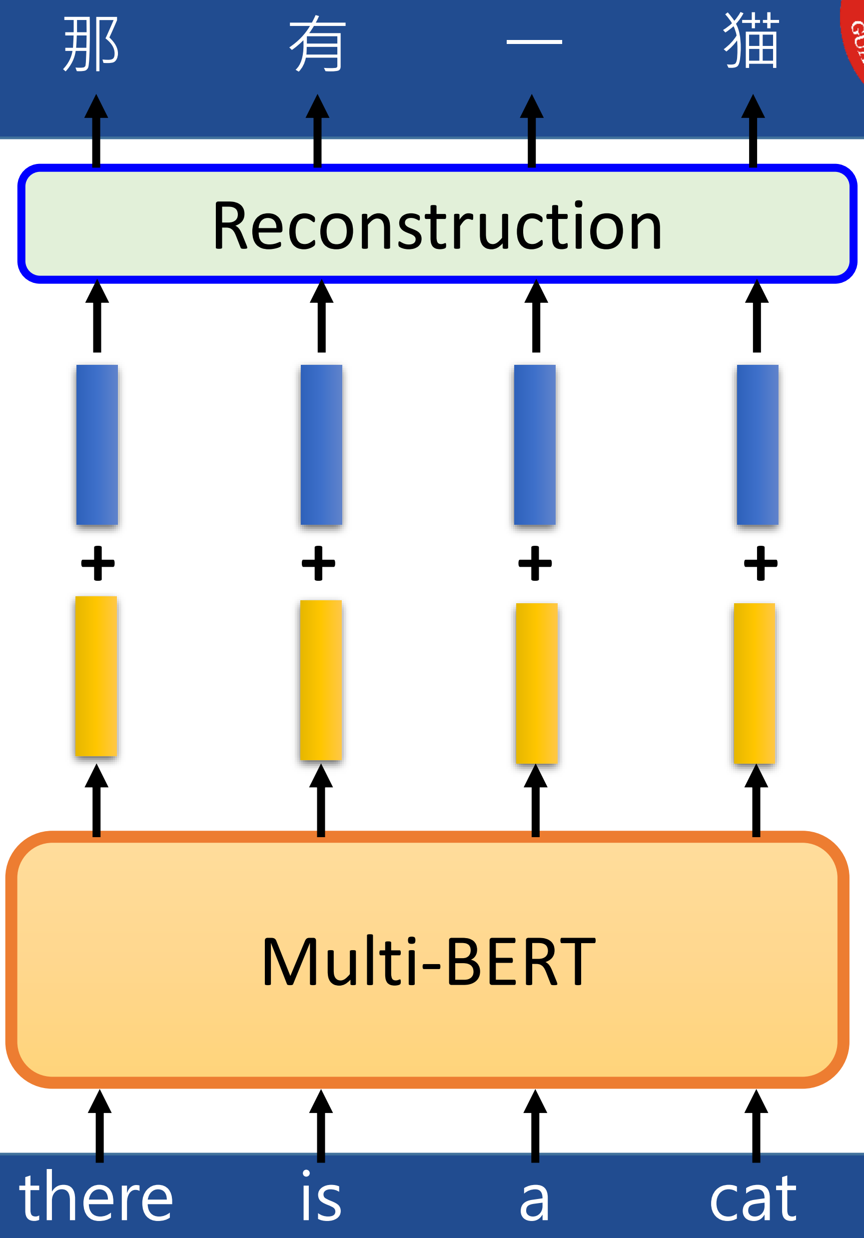
There must be language information.
语言信息



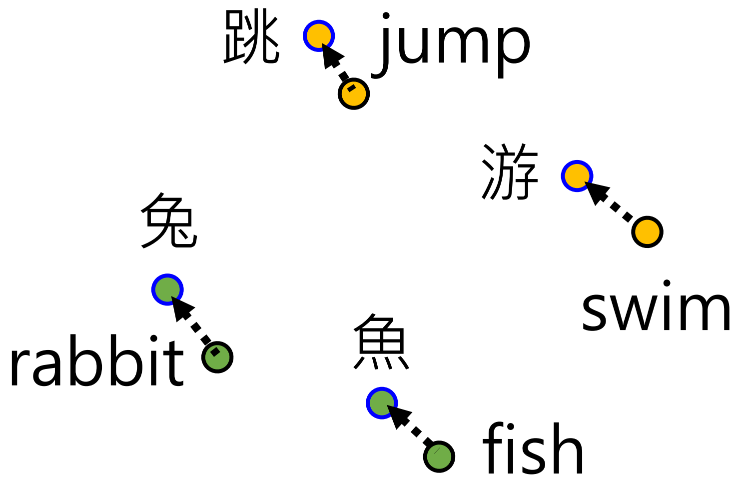
https://arxiv.org/abs/2010.10041

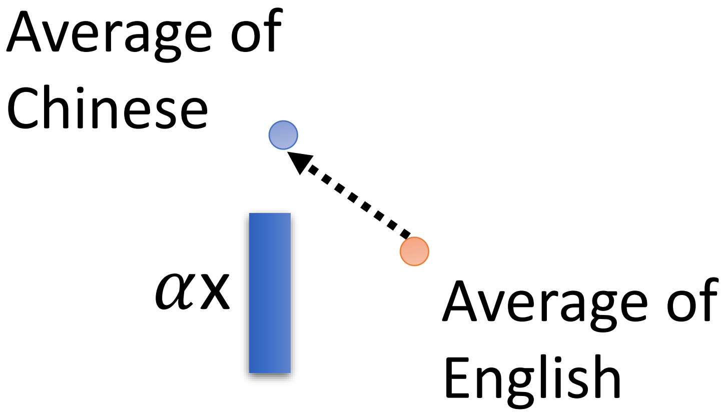
Where is Language?





Where is Language?





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 ☺

Q&A