



07-Self-Supervised Learning (BERT)

[1. Self-supervised Learning](#)

[2. BERT](#)

[2.1 Masking Input](#)

[2.2 Next Sentence Prediction](#)

[2.3 BERT 的實際用途](#)

[2.3.1 GLUE 指標 \(General Language Understanding Evaluation\)](#)

[2.3.2 Case 1 : Sentiment analysis](#)

[2.3.3 Case 2 : POS tagging](#)

[2.3.4 Case 3 : Natural Language Inference \(NLI\)](#)

[2.3.5 Case 4 : Extraction-based Question Answering \(QA\)](#)

[2.4 BERT 訓練難度高](#)

[3. Pre-training a seq2seq model](#)

[4. 為什麼 BERT 有用？](#)

[4.1 Embedding](#)

[4.1.1 相關技術：CBOW](#)

[4.1.2 contextualized embedding](#)

[4.2 Learn More BERT](#)

[5. Multi-lingual BERT](#)

[5.1 Cross-lingual Alignment](#)

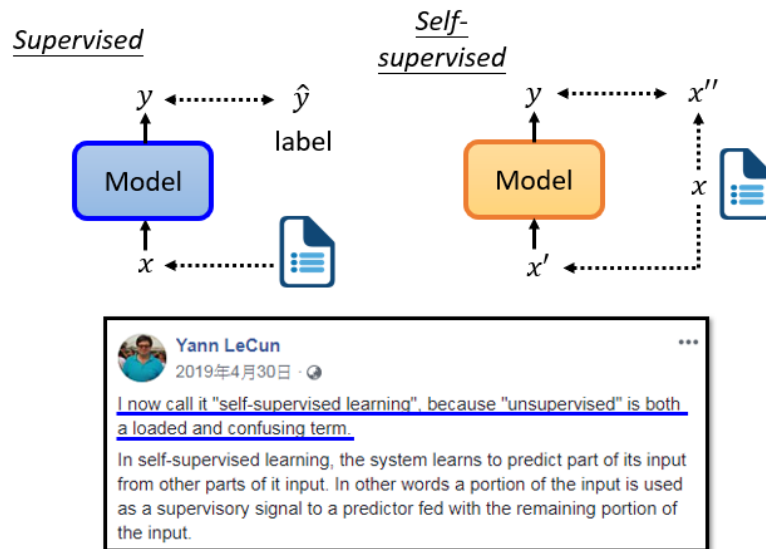
[6. GPT](#)

[6.1 To Learn More](#)

[7. 其他 Self-supervised Learning 應用](#)

1. Self-supervised Learning

self-supervised learning 屬於 **unsupervised learning** 的一種，其資料本身沒有標籤，但是訓練過程中實際上有模型自己生成的標籤



把訓練資料分為兩部分，一部分為**輸入資料**、另一部分為**標註資料**

2. BERT

BERT 是一個 **transformer** 的 **encoder**。BERT 可以輸入一排向量，然後輸出另一排向量，輸出的長度與輸入的長度相同。BERT 一般用於**自然語言處理**，它的輸入是一串文本，也可以輸入語音、圖像等向量序列

訓練 BERT 有兩個任務，分別是 **Masking Input** 及 **Next Sentence Prediction**

2.1 Masking Input

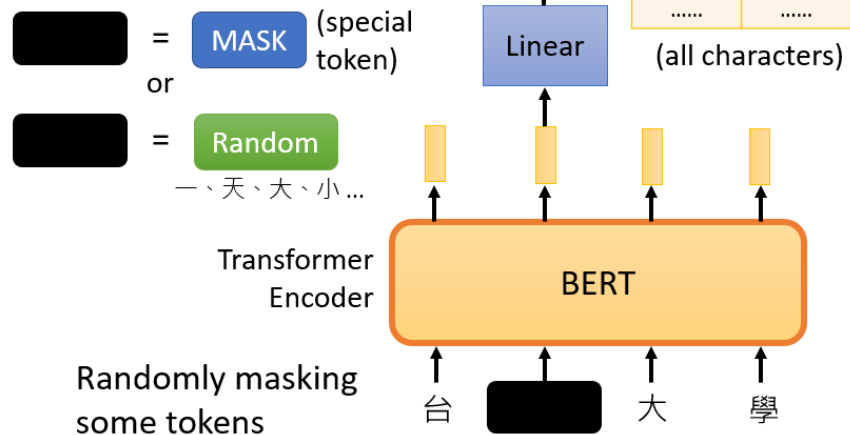
mask 的方法：

- 方法一：用一個特殊的 token "**MASK**" 蓋住句子中的一個詞
- 方法二：隨機把某一個字換成另一個字

兩種方法都可以使用，使用哪種方法也是隨機決定的

Masking Input

<https://arxiv.org/abs/1810.04805>



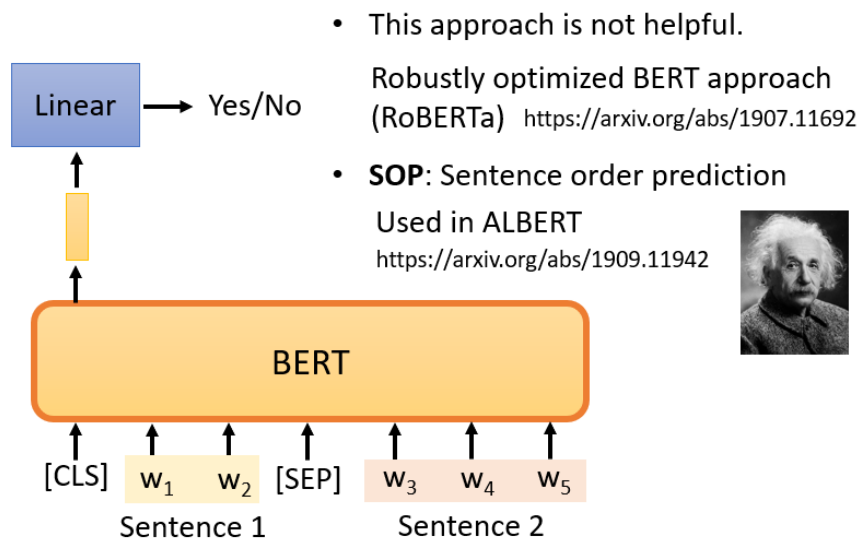
訓練方法：

1. 向 BERT 輸入一個句子，先隨機決定哪一部分的字將被 mask
2. 輸入一個序列，我們把 BERT 的相應輸出看作是另一個序列
3. 在輸入序列中尋找 mask 部分的相應輸出，將這個向量通過一個 linear transform（矩陣相乘），並做 softmax 得到一個分布
4. 用 one-hot vector 表示被 mask 的字符，並使輸出和 one-hot vector 之間的 cross entropy 最小

本質上就是在解決一個分類問題，BERT 要做的是預測什麼字被蓋住

2.2 Next Sentence Prediction

在兩個句子之間添加一個特殊標記 **[SEP]**，代表兩句子的分隔，如此 BERT 就知道是兩個不同的句子，此外還會在句子的開頭添加另一個特殊標記 **[CLS]**



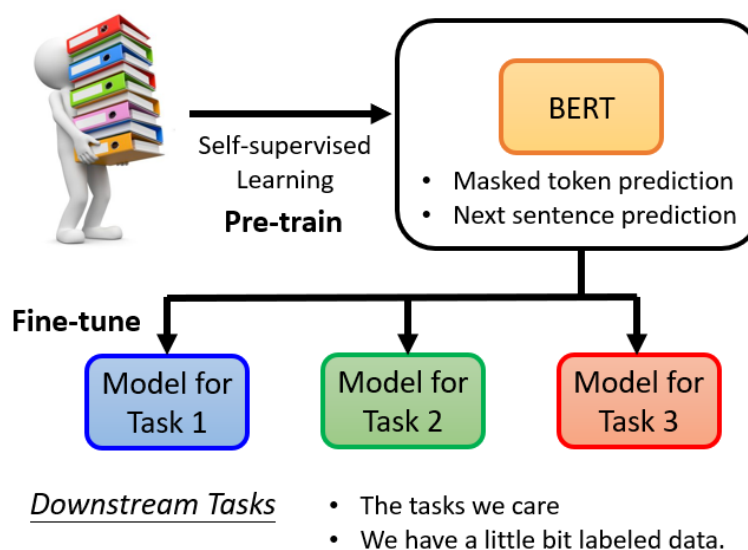
只看 **[CLS]** 的輸出，把它乘以一個 linear transform，做一個二分類問題，輸出 yes/no，判斷兩句是否前後連續

注意：

論文 *Robustly Optimized BERT Approach* (RoBERTa) 指出 Next Sentence Prediction 的方法幾乎沒有幫助，但還有另一種更有用的方法叫做 **Sentence Order Prediction**，該方法選擇兩個句子本來就是連接在一起，但順序可能顛倒或沒有顛倒兩種可能性，BERT 要回答是哪一種可能性。它被用於名為 **ALBERT** 的模型中，該模型是 **BERT** 的進階版本

2.3 BERT 的實際用途

BERT 可以用於其他任務，這些任務不一定與填空有關，它可能是完全不同的東西，這些任務是真正使用 BERT 的任務，其稱為 **downstream tasks**



- 預訓練 (Pre-train)：產生 BERT 的過程
- 微調 (Fine-tune)：利用一些特別的訊息，使 BERT 能夠完成某種任務

透過預訓練及微調讓 BERT 能夠完成各式各樣的 downstream tasks

2.3.1 GLUE 指標 (General Language Understanding Evaluation)

一個測試指標，為了測試 **self-supervised** 學習的能力，通常會在一個任務集上測試它的準確性，取其平均值得到總分

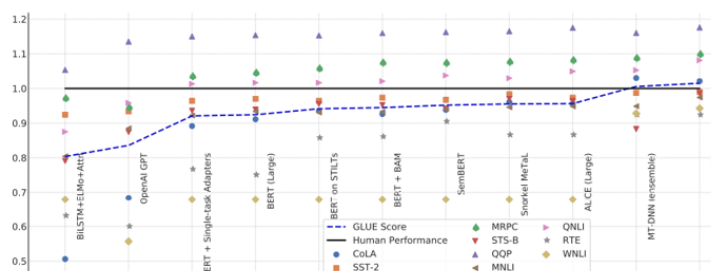
GLUE

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/>)

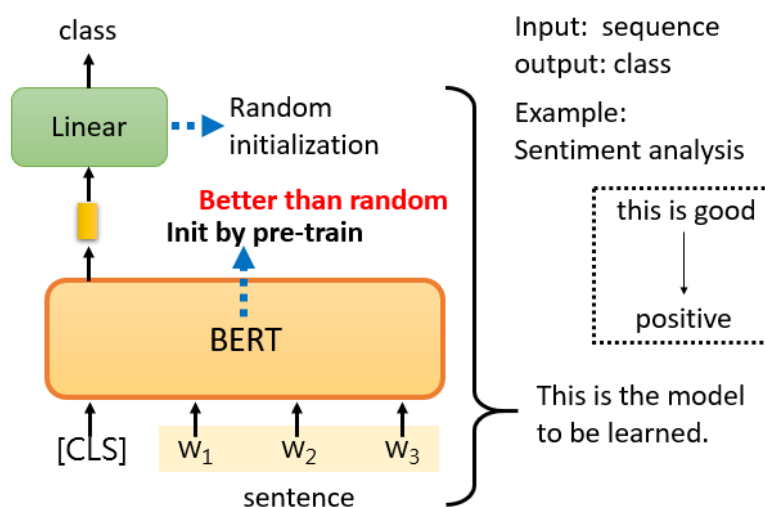
• GLUE scores



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

2.3.2 Case 1 : Sentiment analysis

給 model 一個句子，把 [CLS] 放在句子的前面，只關注 [CLS] 的輸出向量，對它進行 linear transform + softmax，得到類別



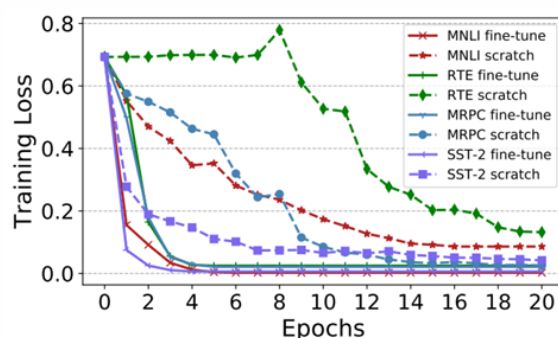
針對情感分析任務作訓練需要一些有標註的資料，訓練的時候，linear transform 和 BERT 模型都是利用 gradient descent 來更新參數的

- linear transform 的參數是隨機初始化的
- BERT 的初始化參數是 pre-train 時學到的參數，此舉會有更好的性能

Pre-train v.s. Random Initialization

(fine-tune)

(scratch)

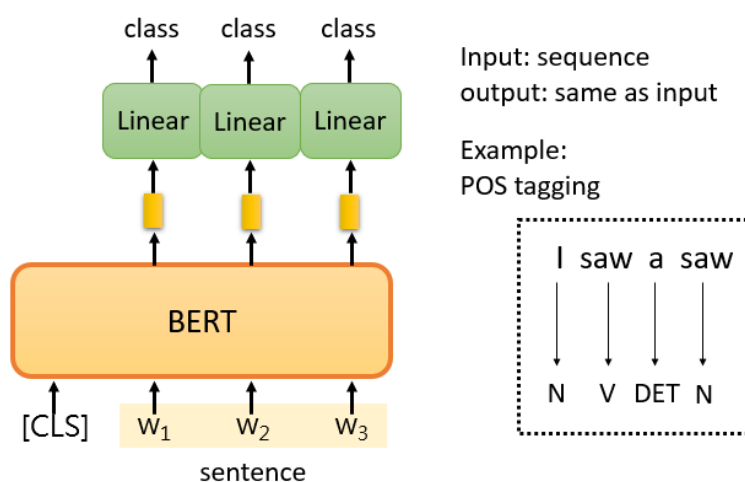


Source of image: <https://arxiv.org/abs/1908.05620>

有 pre-train 的 model 在各個任務上**收斂速度都更快**，在最後也都有**較低的 training loss**

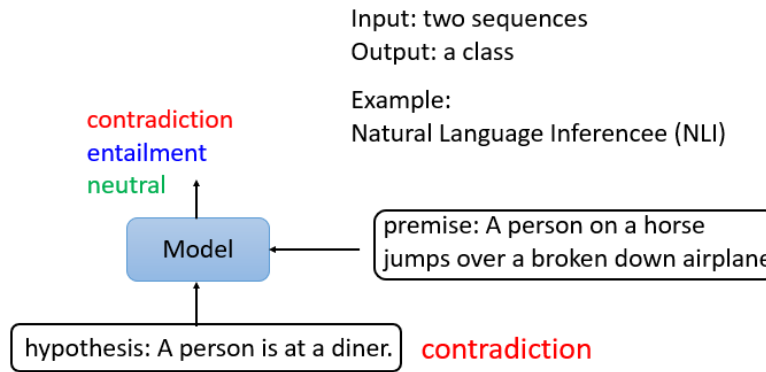
2.3.3 Case 2 : POS tagging

給 model 一個句子，把 [CLS] 放在句子的前面，關注每個字所對應的輸出向量，對每一個向量進行 linear transform + softmax，得到類別

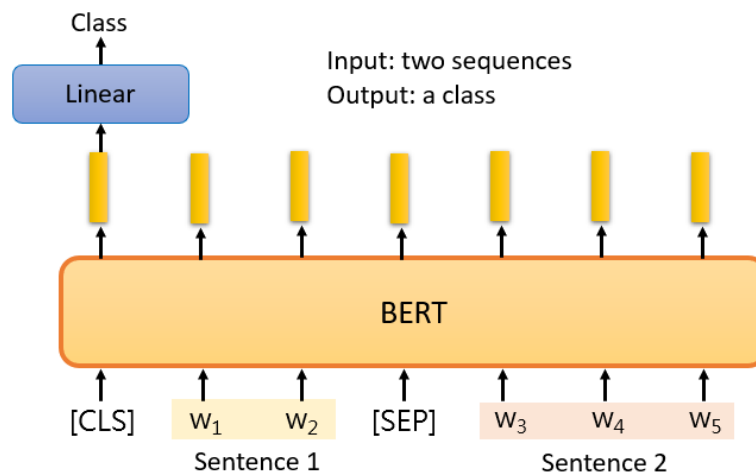


2.3.4 Case 3 : Natural Language Inference (NLI)

給出前提和假設，機器要做的是判斷是否有可能從前提中推斷出假設



給 model 兩個句子，把 [CLS] 放在句子的前面，以 [SEP] 隔開兩個句子，只關注 [CLS] 的輸出向量，對它進行 linear transform + softmax，得到類別



2.3.5 Case 4 : Extraction-based Question Answering (QA)

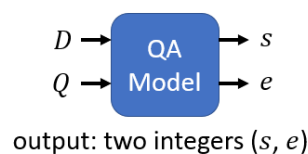
Extraction-based QA 給模型一段文章，要模型回答跟文章相關的問題，保證答案一定在文章裡面

給定文章和問題，輸出兩個整數 s 和 e ，代表文章中的第 s 到 e 個詞彙是模型的答案

• Extraction-based Question Answering (QA)

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

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



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

In meteorology, precipitation is any product of the condensation of atmospheric 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 in scattered locations are called "showers".

What causes precipitation to fall?

gravity

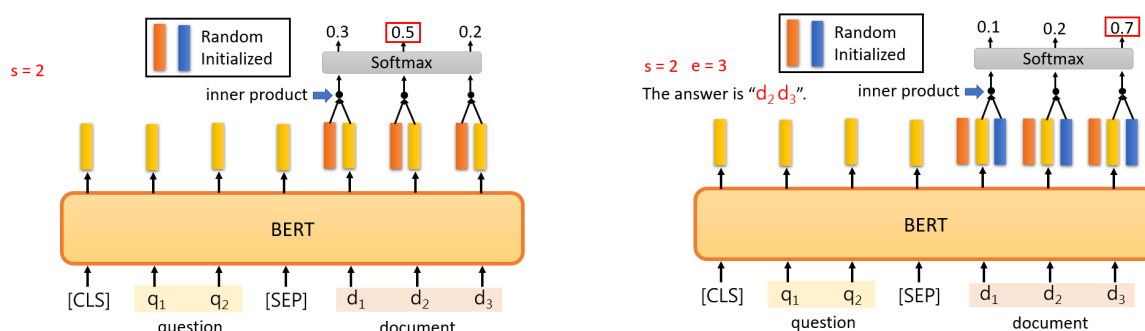
What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?

graupel

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

within a cloud

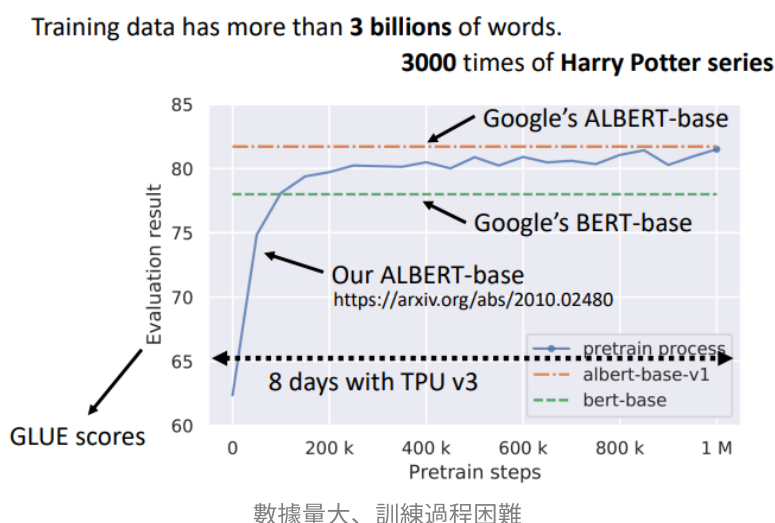
給 model 兩個句子，分別為 question 和 document，把 [CLS] 放在 question 的前面，以 [SEP] 隔開兩個句子，在這個 QA 任務中，只有兩個向量需要隨機初始化，用橘色向量和藍色向量來表示，這兩個向量的長度與 BERT 的輸出相同



方法：

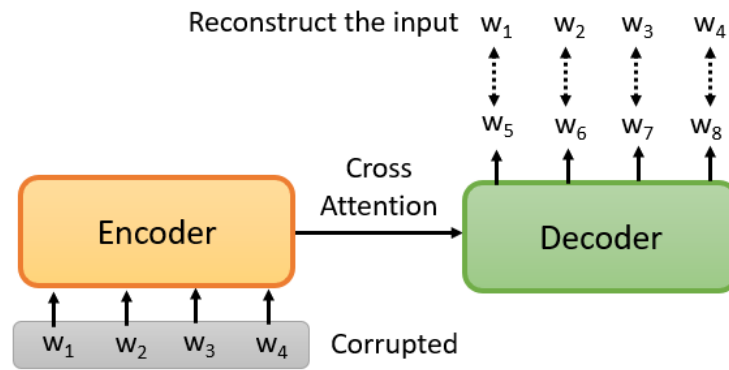
計算 document 中的每一個詞的輸出與橘色向量的 inner product 後，做 softmax 取分數最高的詞彙作為 s ；同樣在計算與藍色向量的 inner product 後，做 softmax 取分數最高的詞彙作為 e

2.4 BERT 訓練難度高



3. Pre-training a seq2seq model

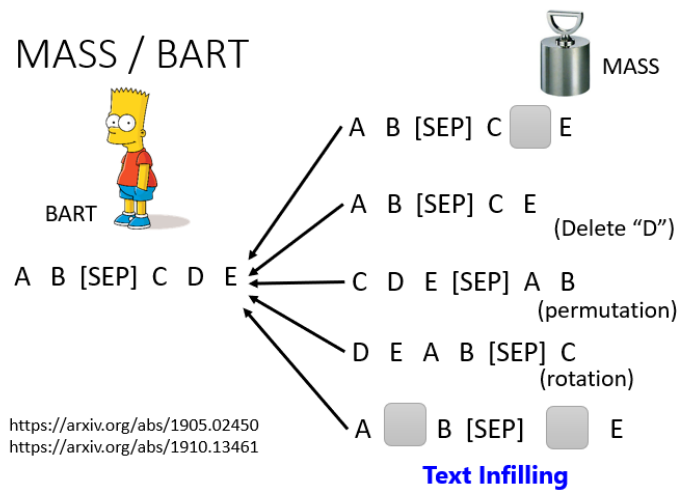
輸入是一串句子，輸出是一串句子，中間用 **cross attention** 連接起來，然後故意在 encoder 的輸入上做些干擾。encoder 看到的是被干擾的結果，decoder 應該輸出句子被破壞前的結果，訓練這個模型實際上是預訓練一個 Seq2Seq 模型



方法：

把某些詞遮住、刪除一些詞，打亂詞的順序，旋轉詞的順序、遮住一些詞再去掉一些詞等等

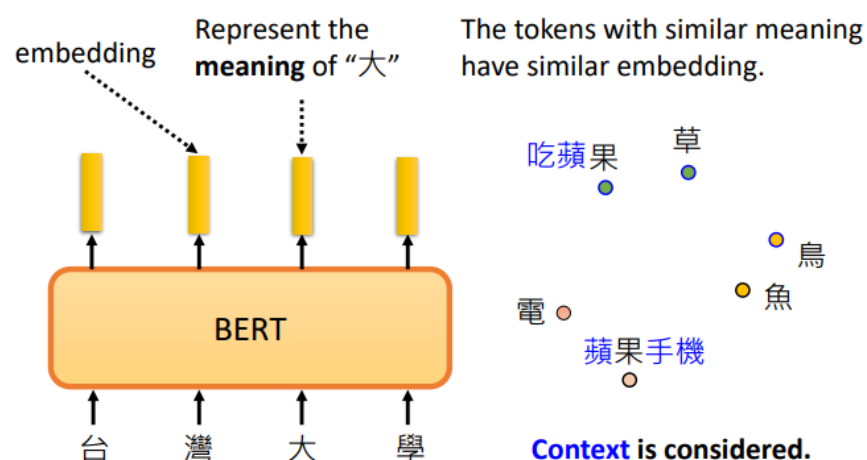
- MASS：把某些詞遮住
- BART：結合全部



4. 為什麼 BERT 有用？

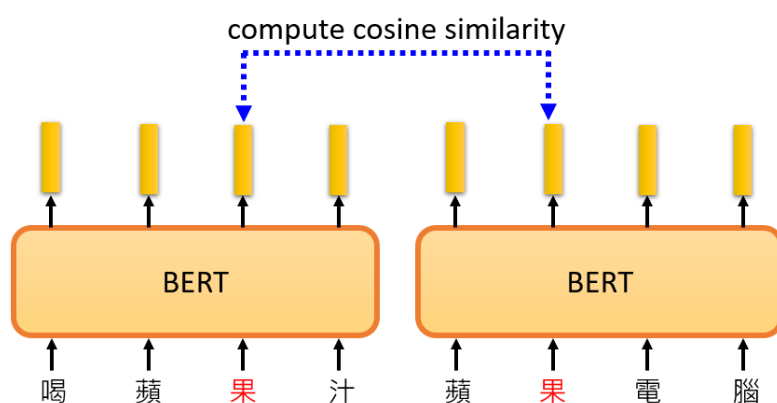
4.1 Embedding

輸入一串文字，每個文字都有對應的向量，稱之為 **embedding**。這些向量代表了輸入詞的含義

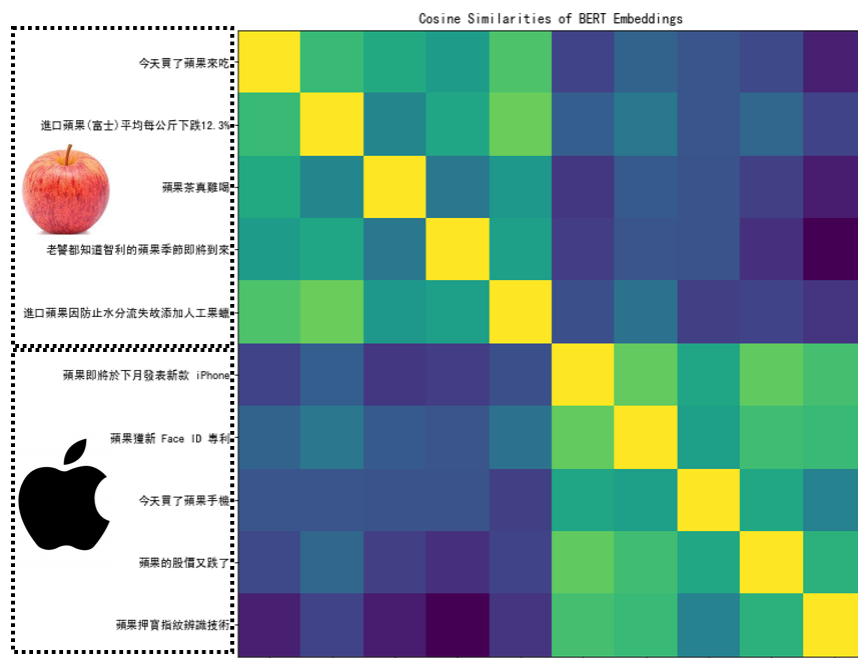


白話地說，把這些詞所對應的向量畫出來，或計算它們之間的距離，可以發現意思比較相似的詞，它們的向量比較接近

訓練 BERT 時，輸入 w_1 、 w_2 、 w_3 和 w_4 ，覆蓋 w_2 要 BERT 預測 w_2 ，而它就是從上下文中提取訊息來預測 w_2 。所以這個向量是其上下文訊息的精華，可以用來預測 w_2 是什麼



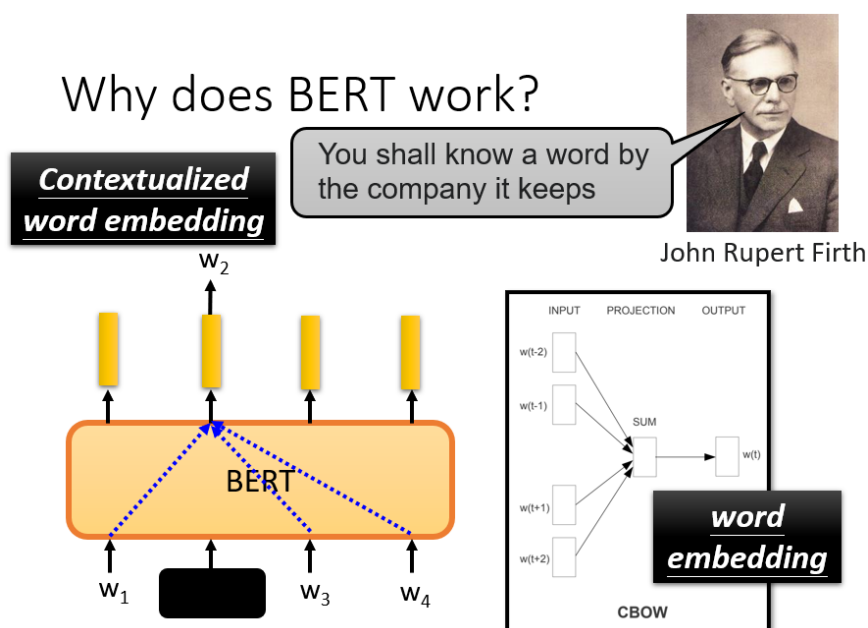
BERT 輸出的這些向量代表了該詞的含義，可以認為 BERT 在填空的過程中已經學會了每個漢字的意思



從上圖針對蘋果的相關性作圖，BERT 知道前五個句子的蘋果代表的是可食用的蘋果，後五個句子的蘋果代表的是蘋果產品

4.1.1 相關技術：CBOW

CBOW 所做的，與 BERT 一樣。做一個空白，並要求它預測空白處的內容。由於算力原因，CBOW 是一個非常簡單的模型，只使用了兩個變換。今天的 BERT，就相當於一個深度版本的 CBOW



4.1.2 contextualized embedding

BERT 還可以根據不同的上下文從相同的詞匯中產生不同的嵌入，因為它是詞嵌入的高級版本，考慮了上下文，BERT 抽取的這些向量或嵌入也稱為 **contextualized word embedding**

4.2 Learn More BERT

To Learn More

BERT (Part 1)



https://youtu.be/1_gRK9EIQpc

BERT (Part 2)

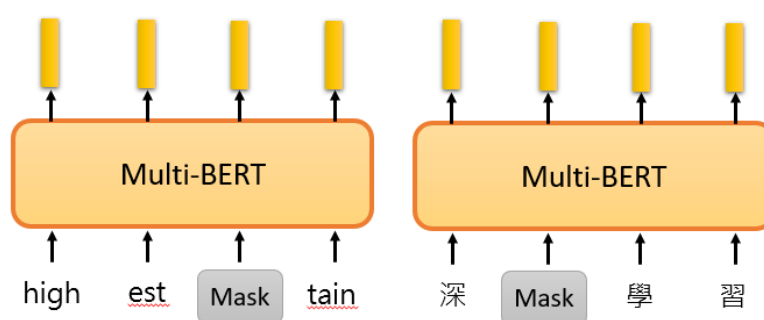


<https://youtu.be/Bywo7m6ySlk>

https://youtu.be/1_gRK9EIQpc、<https://youtu.be/Bywo7m6ySlk>

5. Multi-lingual BERT

使用多語言來進行預訓練，比如中文、英文、德文、法文等等的填空題來訓練 BERT，稱為 **multi-lingual BERT**



Training a BERT model by many different languages.

Google 訓練了一個 multi-lingual BERT，做了 104 種語言的填空題 pre-train。神奇之處是如果用英文問答數據做 fine-tune，但是測試中文問答，BERT 也可以表現得很好

- English: SQuAD, Chinese: DRCD

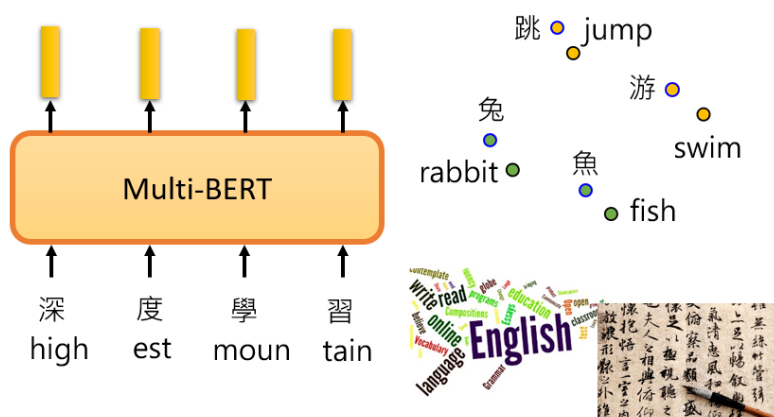
| Model | Pre-train | Fine-tune | Test | EM | F1 |
|-------|---------------|-------------------|---------|------|------|
| QANet | none | Chinese | Chinese | 66.1 | 78.1 |
| BERT | Chinese | Chinese | | 82.0 | 89.1 |
| | 104 languages | Chinese | | 81.2 | 88.7 |
| | | English | | 63.3 | 78.8 |
| | |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

F1 score of Human performance is 93.30%

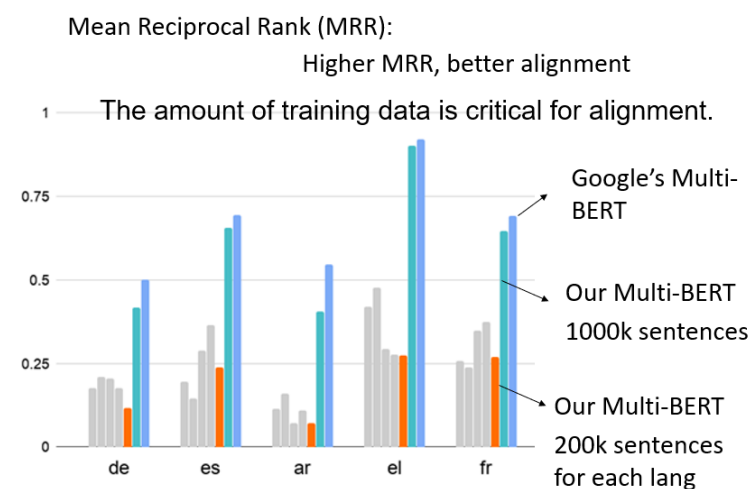
This work is done by 劉記良、許宗嫻
<https://arxiv.org/abs/1909.09587>

5.1 Cross-lingual Alignment

一個簡單的解釋上述現象，也許對於 multi-lingual 的 BERT 來說，不同的語言並沒有那麼大的差異。無論用中文還是英文顯示，對於具有相同含義的單詞，它們的 **embedding** 都很接近。中文的跳與英文的 jump 接近，中文的魚與英文的 fish 接近，也許在學習過程中 BERT 已經自動學會了



MRR (Mean Reciprocal Rank) 的值越高，同樣意思不同語言的詞其向量越接近



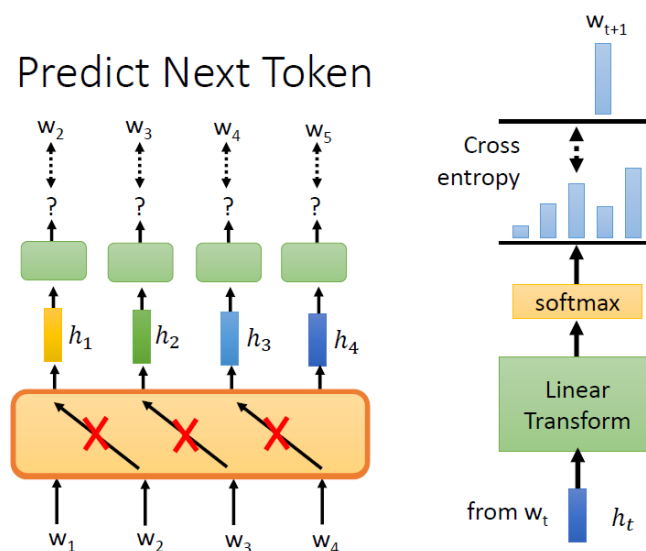
<https://arxiv.org/abs/2010.10938>

投影片來源: 許宗嫻同學碩士口試投影片

一開始使用的資料較少，每種語言只使用了 20 萬個句子，訓練的模型的結果並不好。之後增加到 100 萬，有了更多的數據，**BERT 可以學習 alignment**，所以資料量是不同語言能否成功對齊的一個非常關鍵的因素

6. GPT

架構類似 transformer encoder，用已知的詞來預測接下來的詞



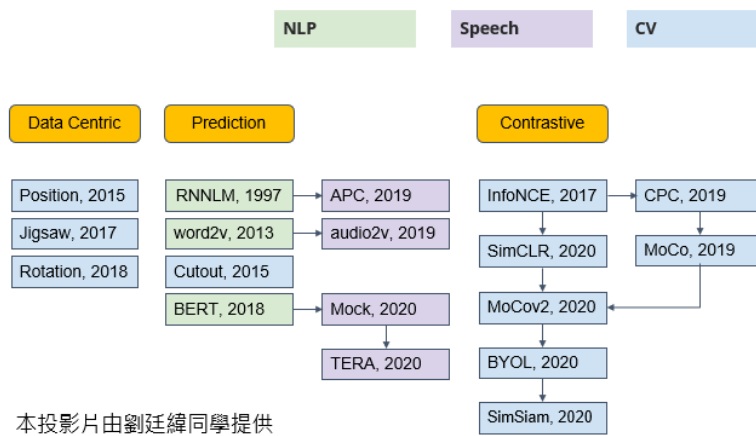
6.1 To Learn More



<https://youtu.be/DOG1L9lvsDY>

<https://youtu.be/DOG1L9lvsDY>

7. 其他 Self-supervised Learning 應用



本投影片由劉廷緯同學提供