自然語言處理與應用

Natural Language Processing and Applications

BERT and its Family

Instructor: 林英嘉 (Ying-Jia Lin)

2025/03/31

Course GitHub

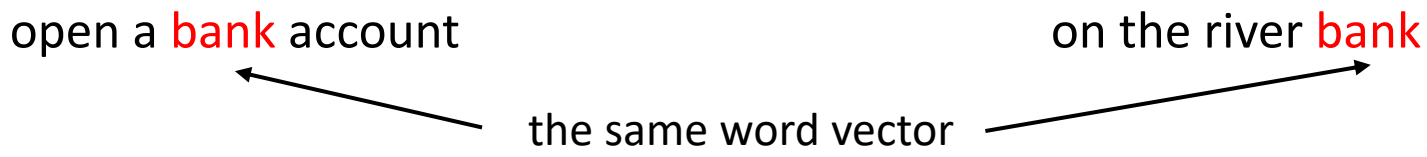
Slido # NLP_0331

Outline

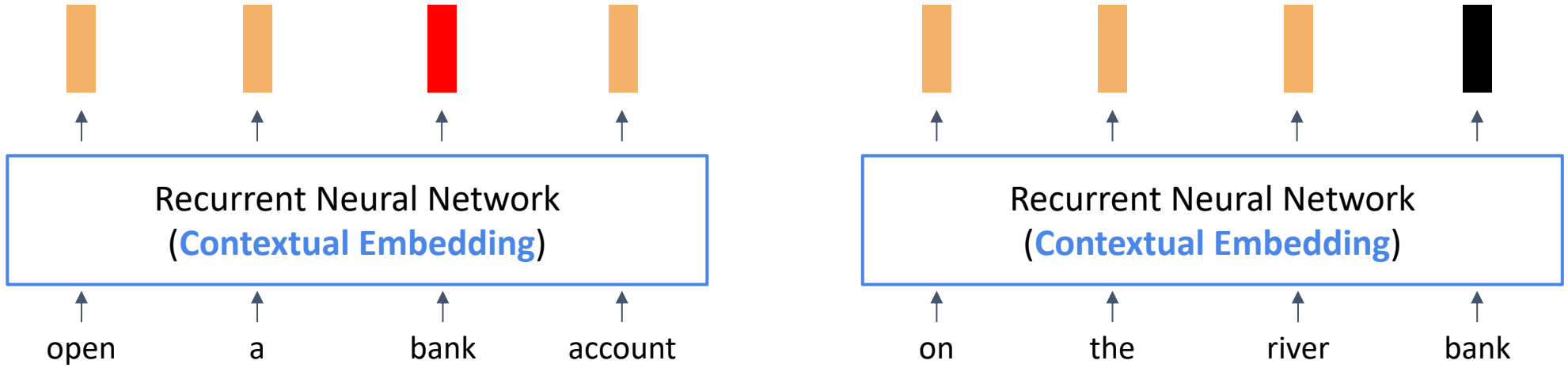
- ELMo (20 min)
- BERT and its Family [50 min]
- PyTorch (last time) [20 min]
- Announcement [10 min]
- Decoding [30 min]
- Quiz [20 min]

Static Embedding vs. Contextual Embedding

Static Embedding (traditional word embedding)

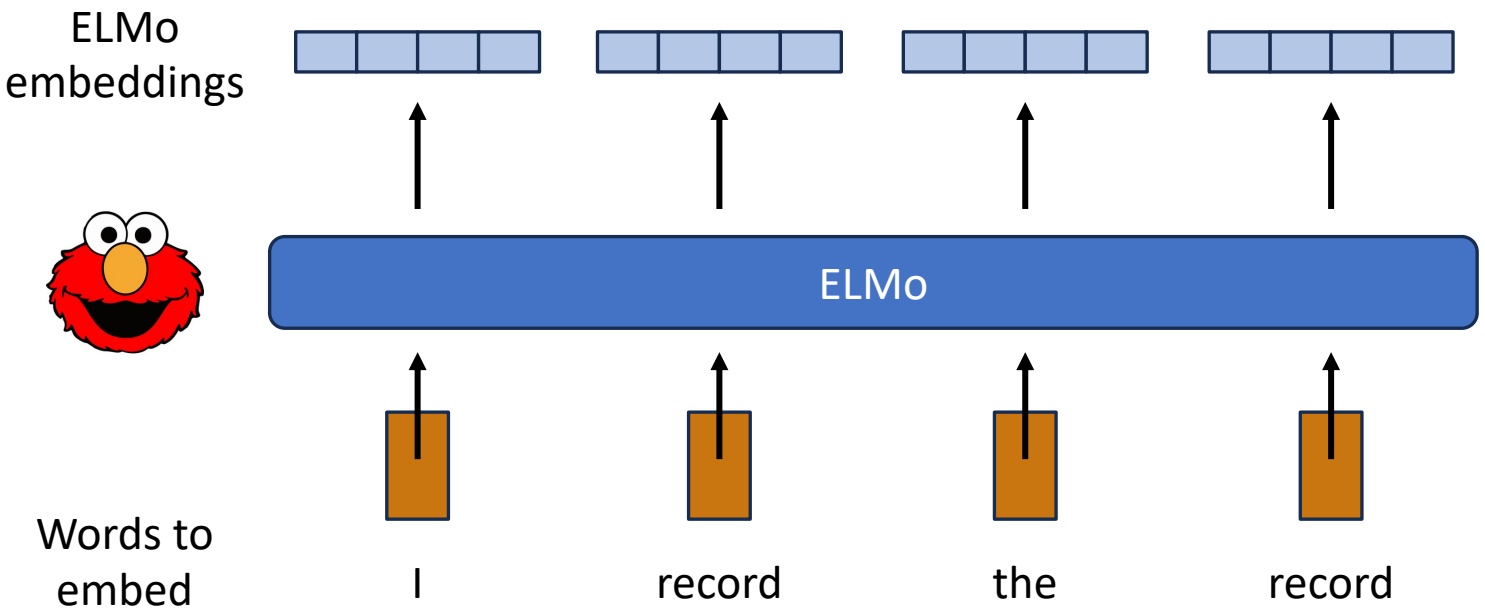


Contextual Embedding (會隨著時間點而改變的 word embedding)



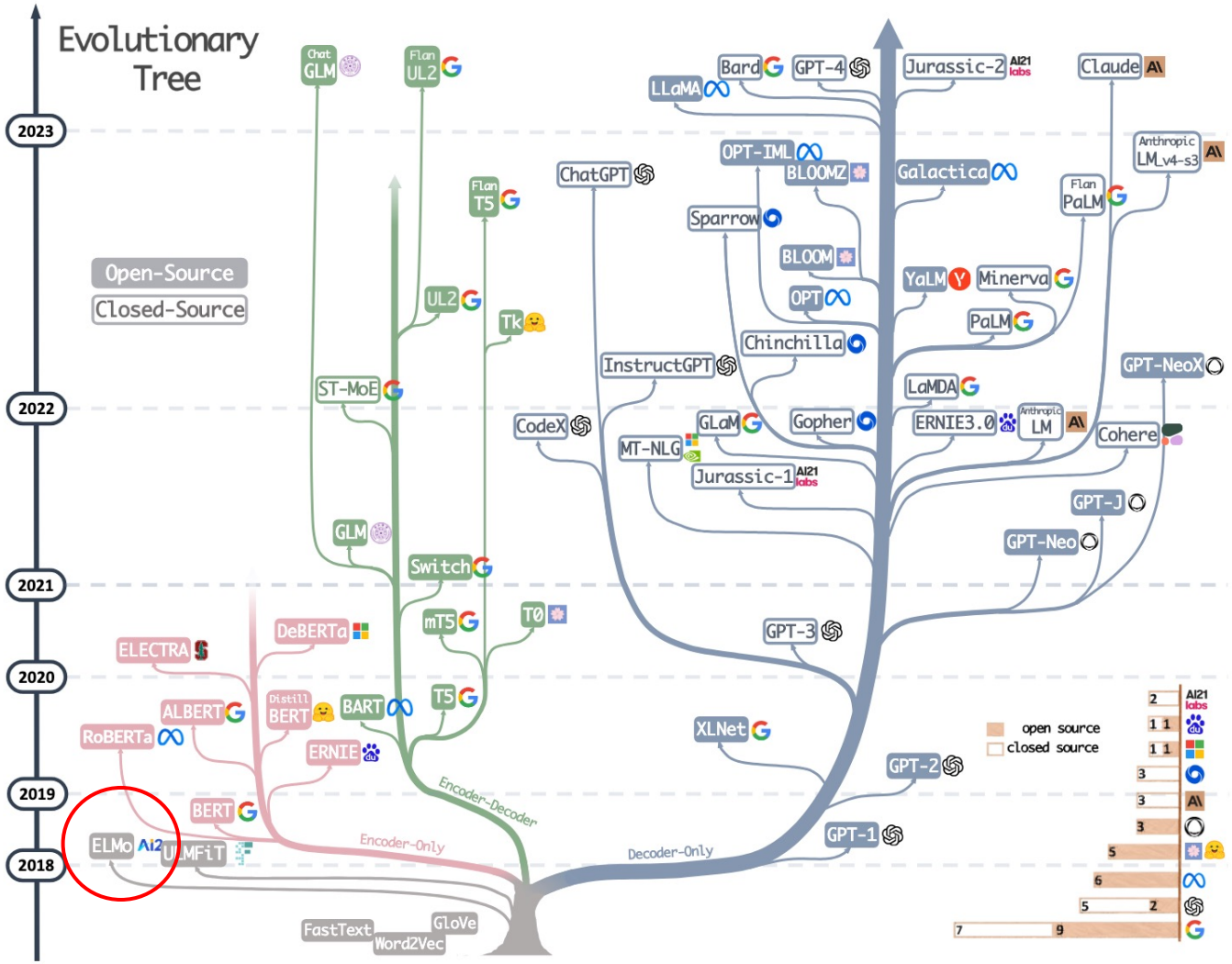
Context Matters: ELMo (Peters et al., 2018)

Rather than using a fixed representation for each word, ELMo considers the entire sentence before assigning each word in it an embedding.



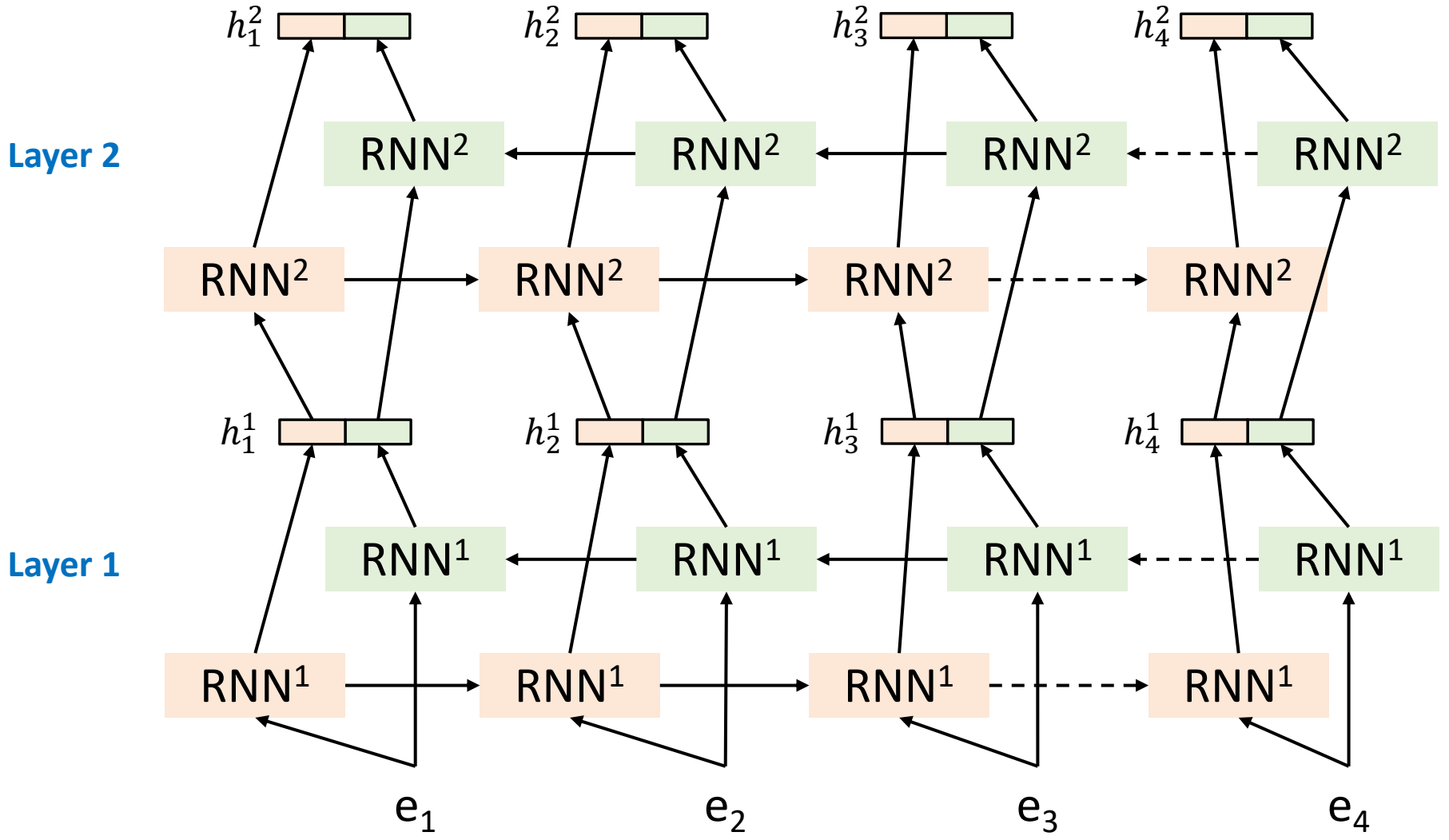
Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. Deep Contextualized Word Representations. NAACL 2018.

Evolutionary Tree of Large Language Models



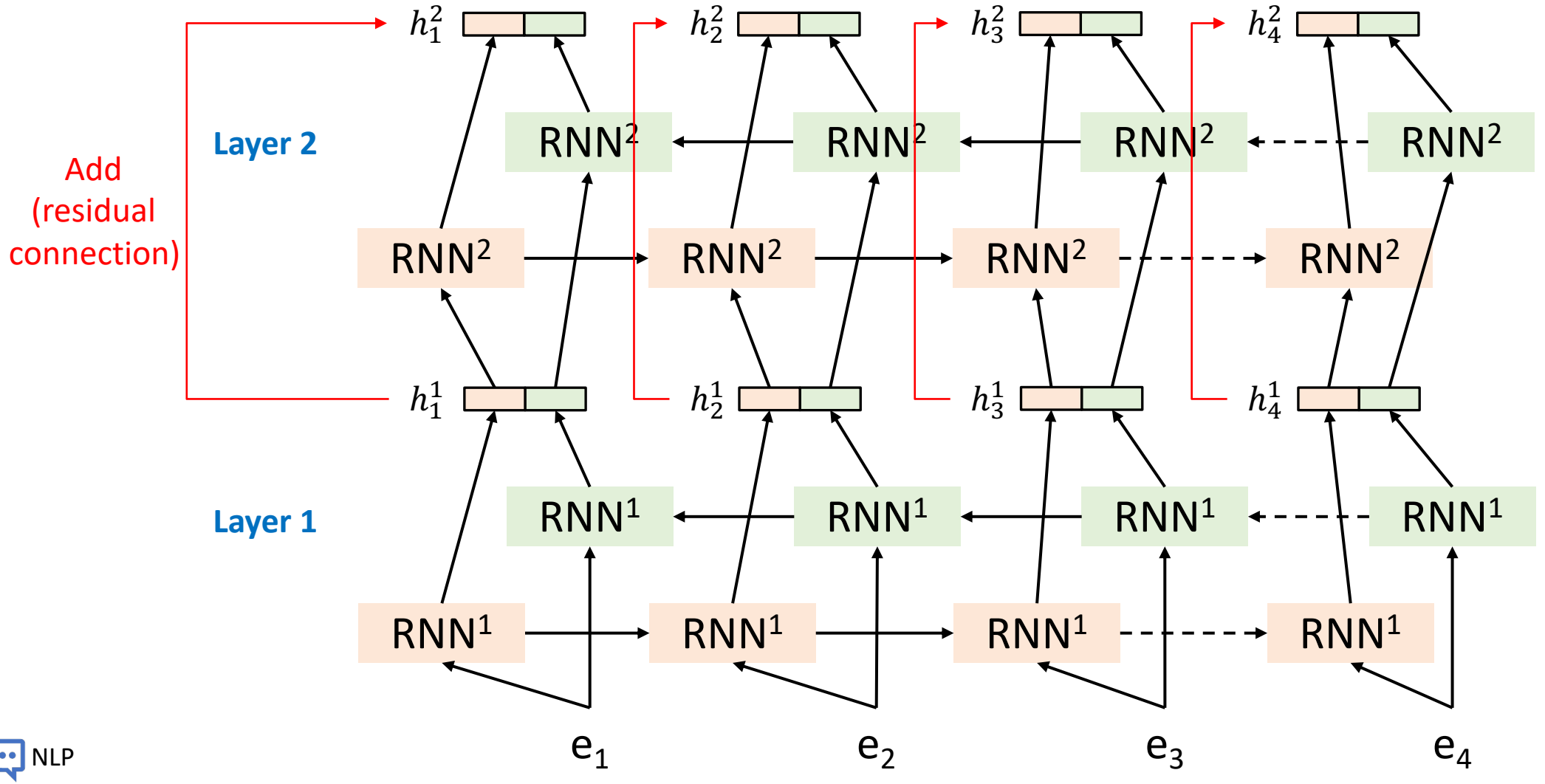
Yang, Jingfeng, et al. "Harnessing the power of llms in practice: A survey on chatgpt and beyond." ACM Transactions on Knowledge Discovery from Data 18.6 (2024): 1-32.

Bidirectional RNN with 2 Layers

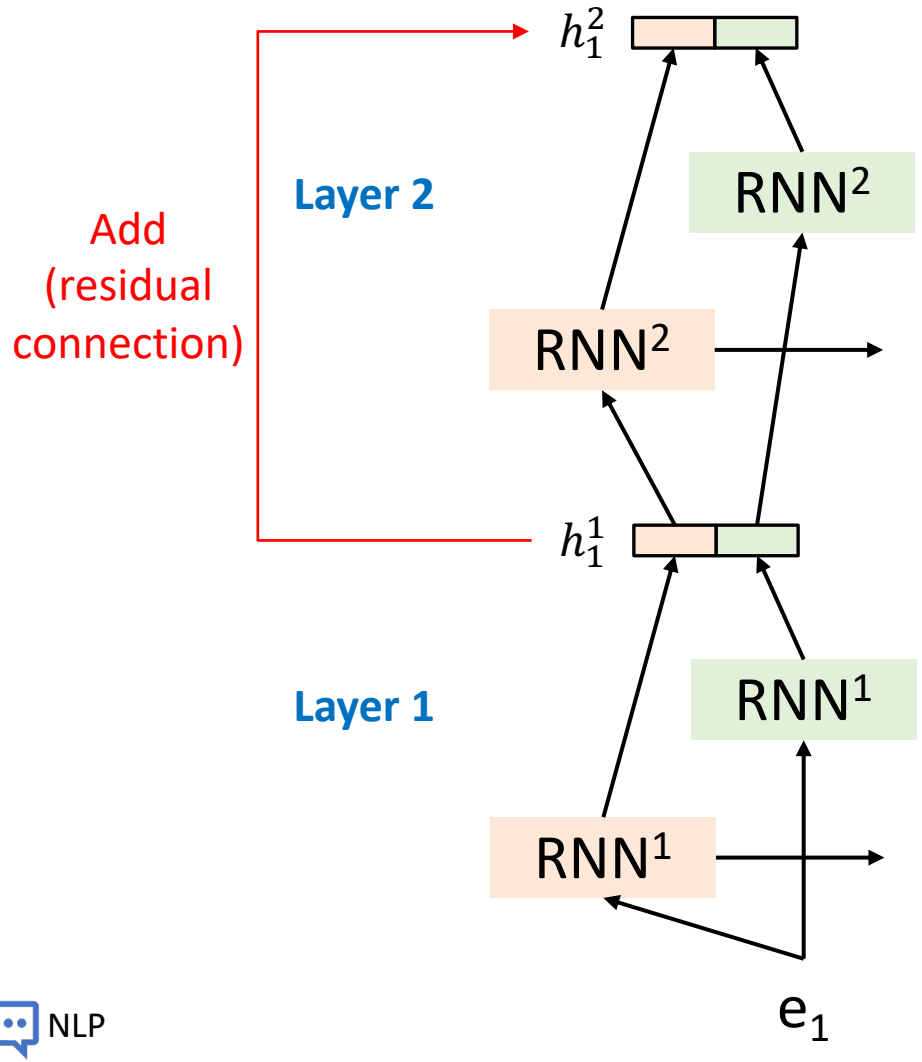


ELMo (1)

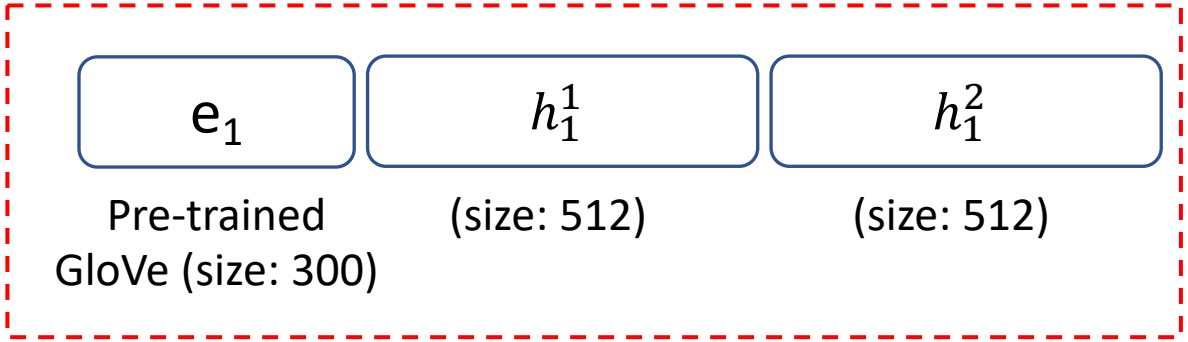
- ELMo 是一個兩層的Bi-LSTM
- 在第1層和第2層之間有: residual connections



ELMo (2)

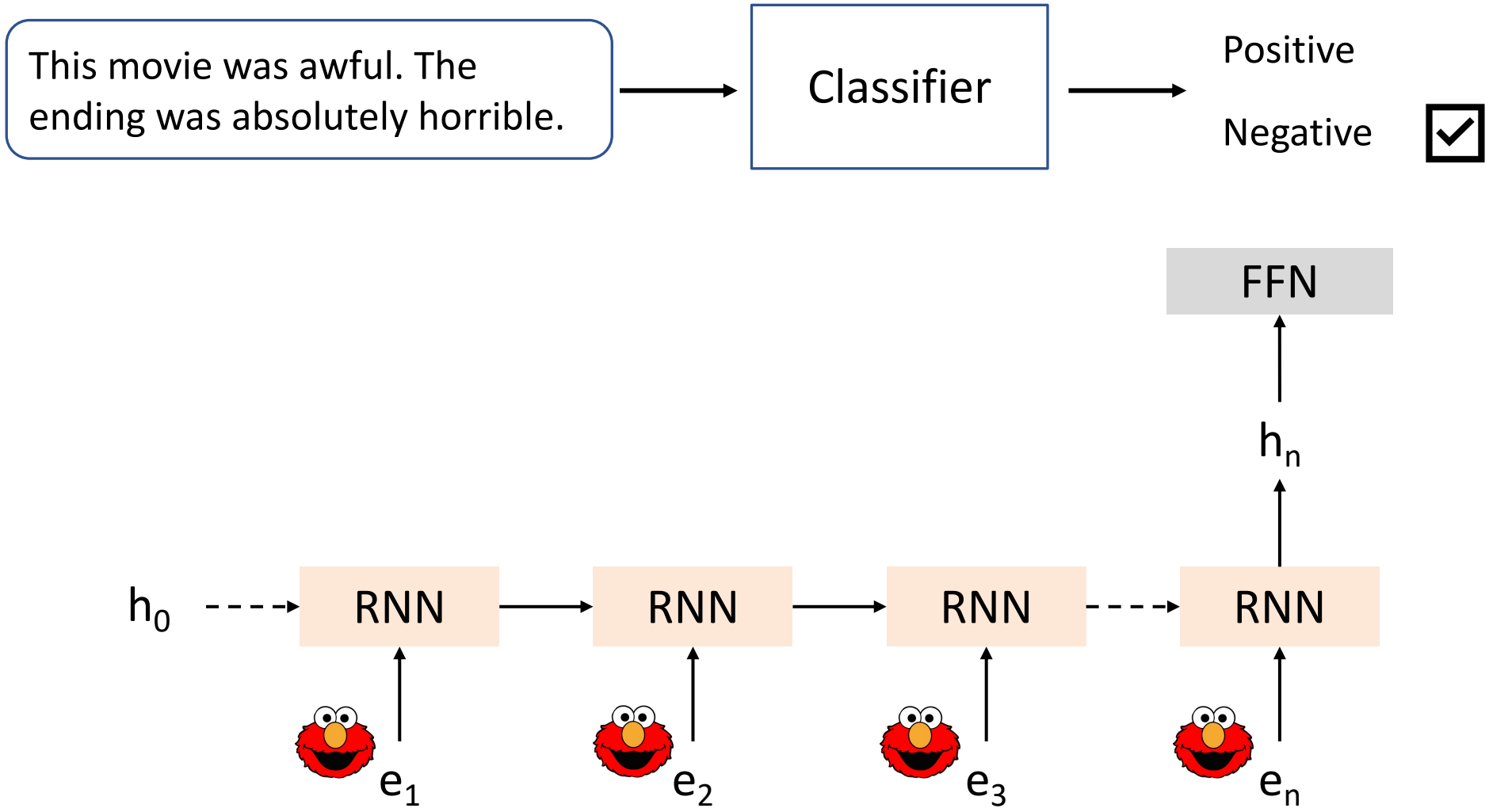


- ELMo 的輸出是 embedding
- ELMo 是 contextual embedding
  - 每個 token 的 embedding 由 input embedding 和每一層的 hidden state 做 concatenation



token$_1$ 的 token embedding (size: 1324)

ELMo 的使用方法


This movie was awful. The ending was absolutely horrible.
Classifier
Positive
Negative
FFN
h_n
h_0
RNN
RNN
RNN
RNN
e_1
e_2
e_3
e_n

ELMo 的 Deep 可能不是 Deep Layers

- ELMo 的 Deep 可能表達的是深度的語意資訊

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular play on Alusik 's grounder {...}	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {...}	{...} they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

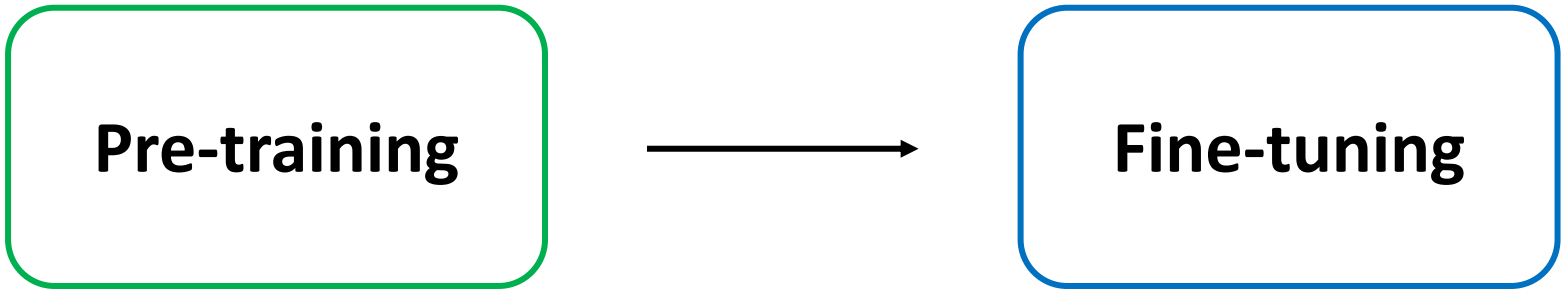
跟球類有關的 play

跟表演有關的 play

Table 4: Nearest neighbors to "play" using GloVe and the context embeddings from a biLM.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. Deep Contextualized Word Representations. NAACL 2018.

The Pre-training then Fine-tuning Paradigm

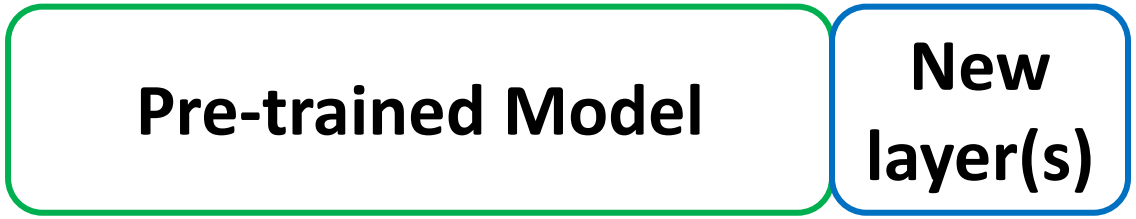


在大量資料上進行訓練，通常是自監督式 (Self-Supervised Training)

在目標資料上 (Down-stream tasks) 進行訓練，通常是監督式 (Supervised Training)，也就是需要有標註的資料才能進行模型訓練

Feature-based and Fine-tuning approaches

- Feature-based approach – ELMo
  - ELMo 並無針對下游任務設計模型結構
  - ELMo 是一種外掛、增幅器，接在別人的工作上
- Fine-tuning approach – GPT, BERT
  - 涵蓋下游任務設計



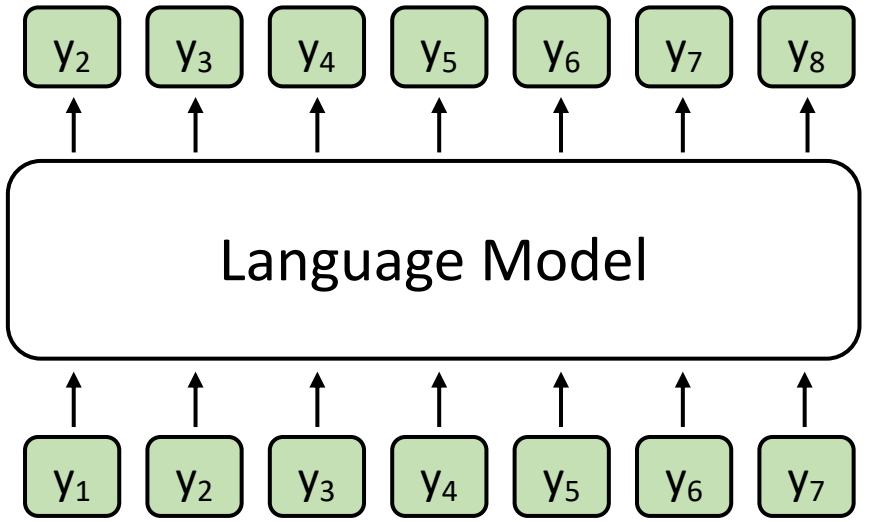
Static Embedding vs. Contextual Embedding

Type	Models
Static Embedding	Word2Vec, GloVe, …
Contextual Embedding	ELMo, BERT, GPT, …

GPT-1

Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018).

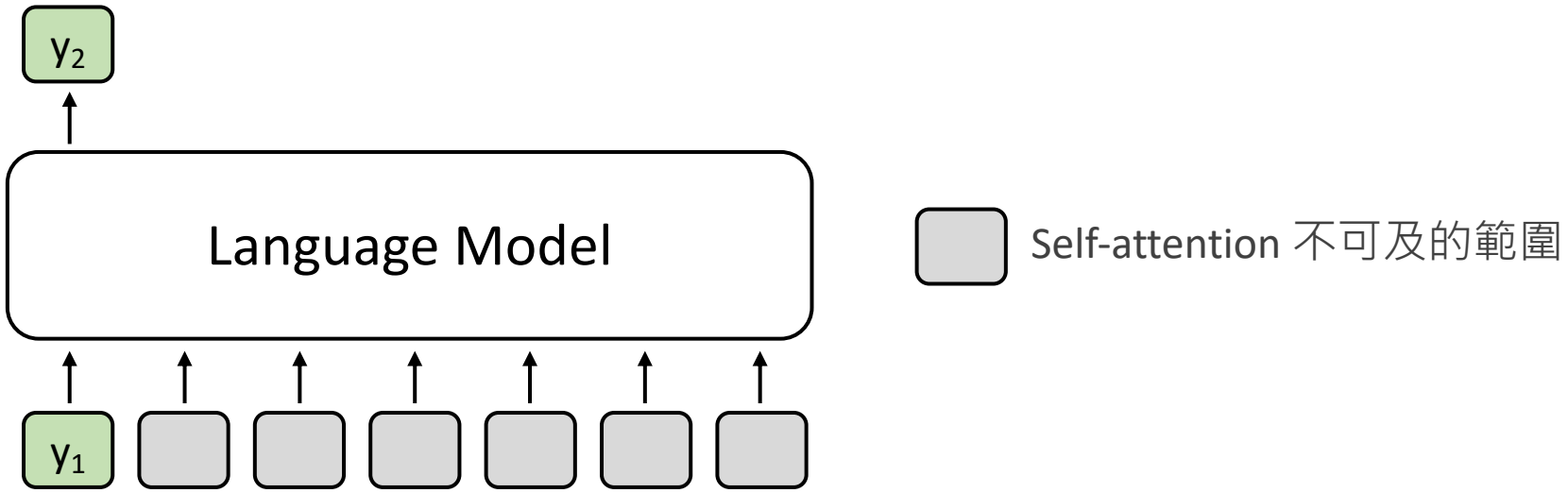
- 只用 Transformer decoder layers
- 訓練模型最大化每個時間點的機率：$P(y_t|y_1, y_2, \ldots, y_{t-1})$



GPT-1

Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018).

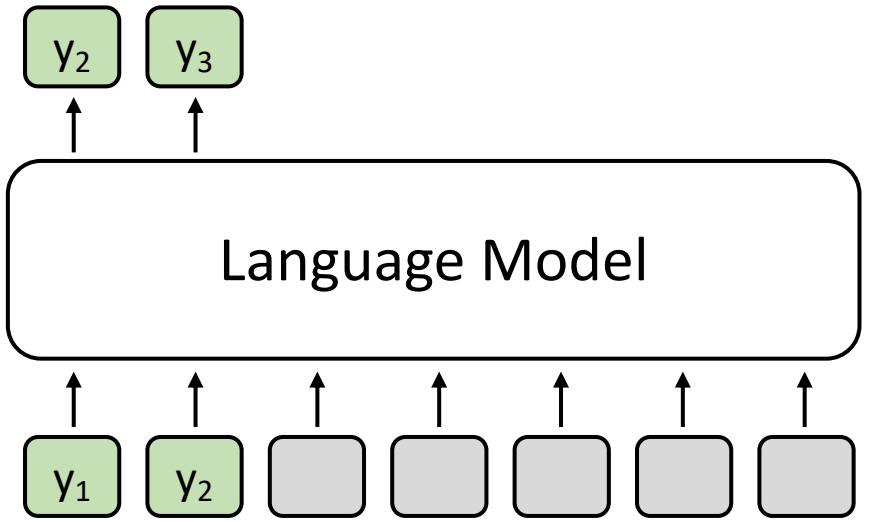
- 只用 Transformer decoder layers
- 訓練模型最大化每個時間點的機率：$P(y_t|y_1, y_2, \ldots, y_{t-1})$



GPT-1

Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018).

- 只用 Transformer decoder layers
- 訓練模型最大化每個時間點的機率：$P(y_t|y_1, y_2, \ldots, y_{t-1})$



Self-attention 不可及的範圍

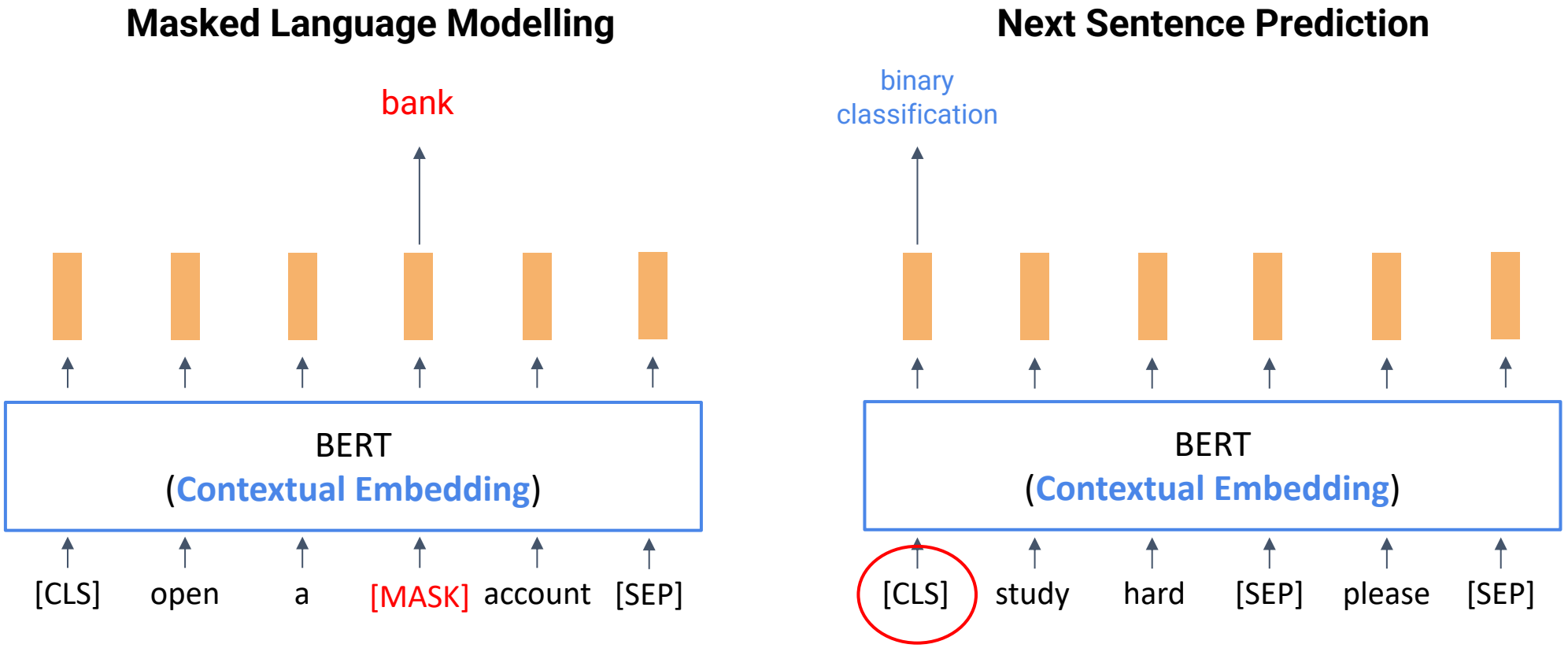
代表 GPT 雖然用了 Transformer，但還是單向

GPT Family

Model	Pre-training method	Parameters	Dataset	Release
GPT-1	Transformer decoder followed by linear-softmax	117 M	BooksCorpus 4.5 GB	2018.06.11
GPT-2	Same as GPT-1 but different normalized layer	1.5 B	WebText 40 GB	2019.02.14
GPT-3	Larger version of GPT-2	175 B	CommonCrawl 45 TB	2020.05.15
InstructGPT	Language modeling + Additional fine-tuning	175 B	Undisclosed	2022.05.04

BERT (Bidirectional Encoder Representations from Transformers)

- BERT 有兩種預訓練任務：



Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." NAACL 2019.

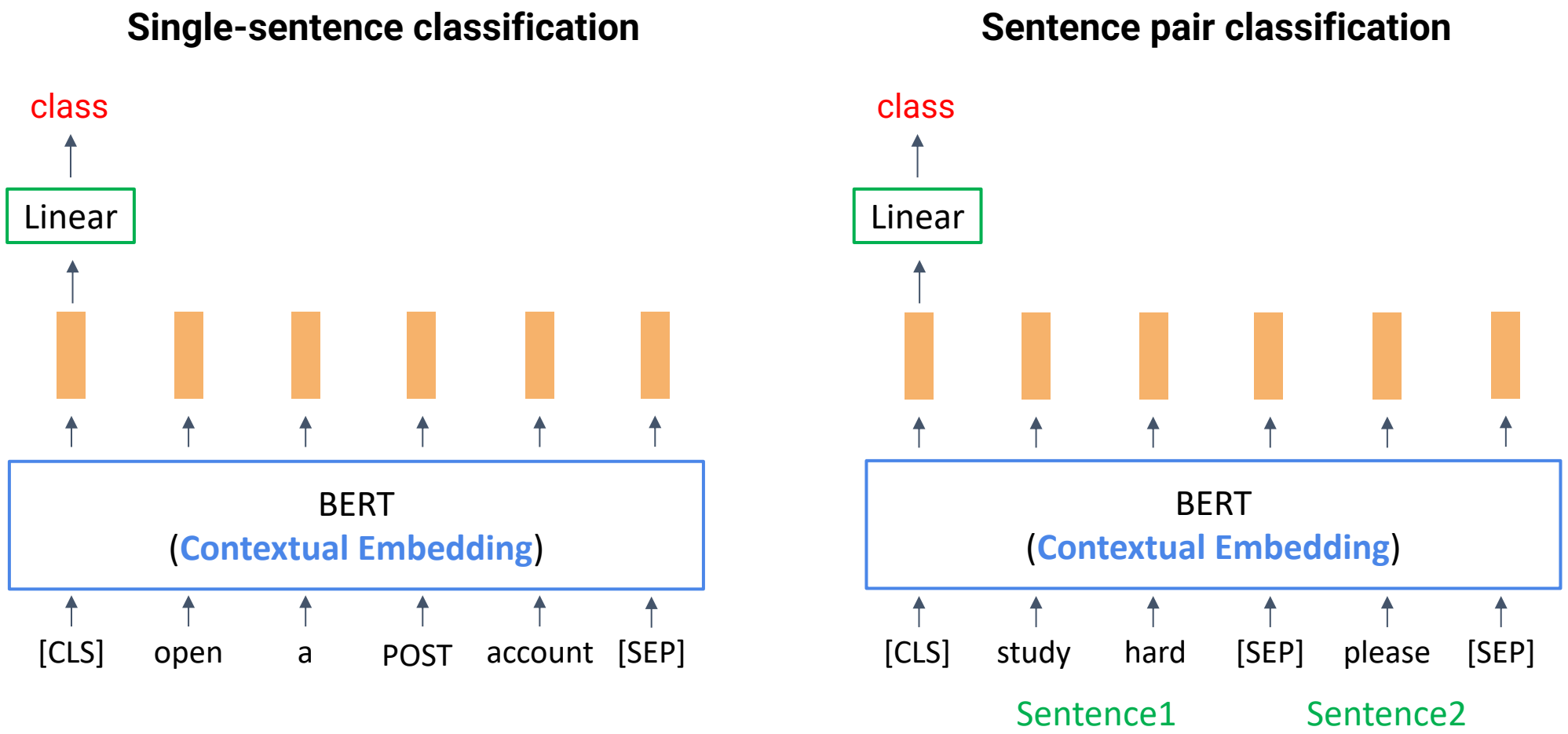
Next Sentence Prediction (NSP)

- Binarized next sentence prediction task
- Aimed at making BERT better on understanding the relationship between two sentences

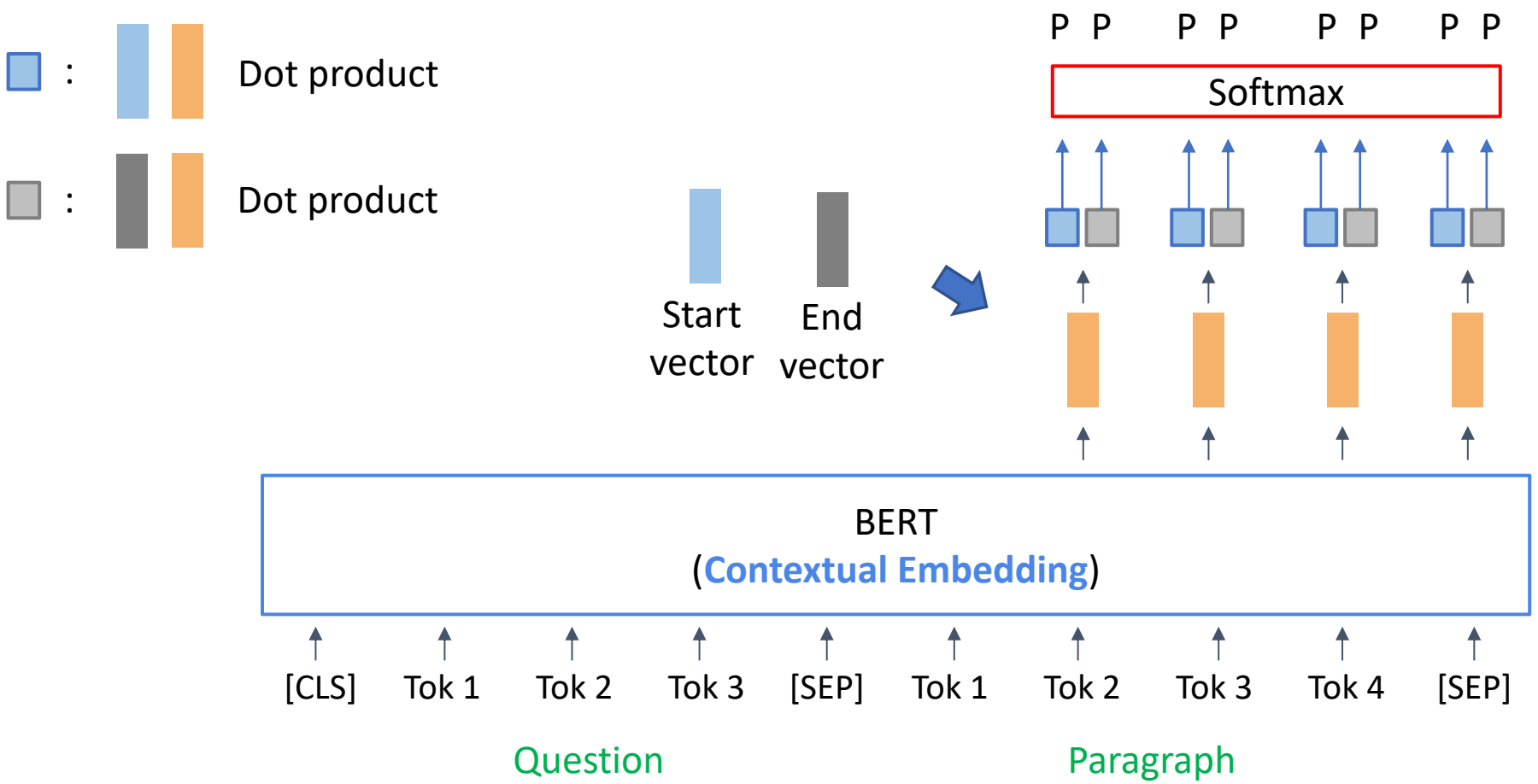
$$\text{Input} = \texttt{[CLS] the man went to [MASK] store [SEP]}$$
$$\texttt{he bought a gallon [MASK] milk [SEP]}$$
$$\text{Label} = \texttt{IsNext}$$

$$\text{Input} = \texttt{[CLS] the man [MASK] to the store [SEP]}$$
$$\texttt{penguin [MASK] are flight ##less birds [SEP]}$$
$$\text{Label} = \texttt{NotNext}$$

Fine-tuning BERT (Sentence Classification)



Fine-tuning BERT (Question Answering)



Question Answering Example

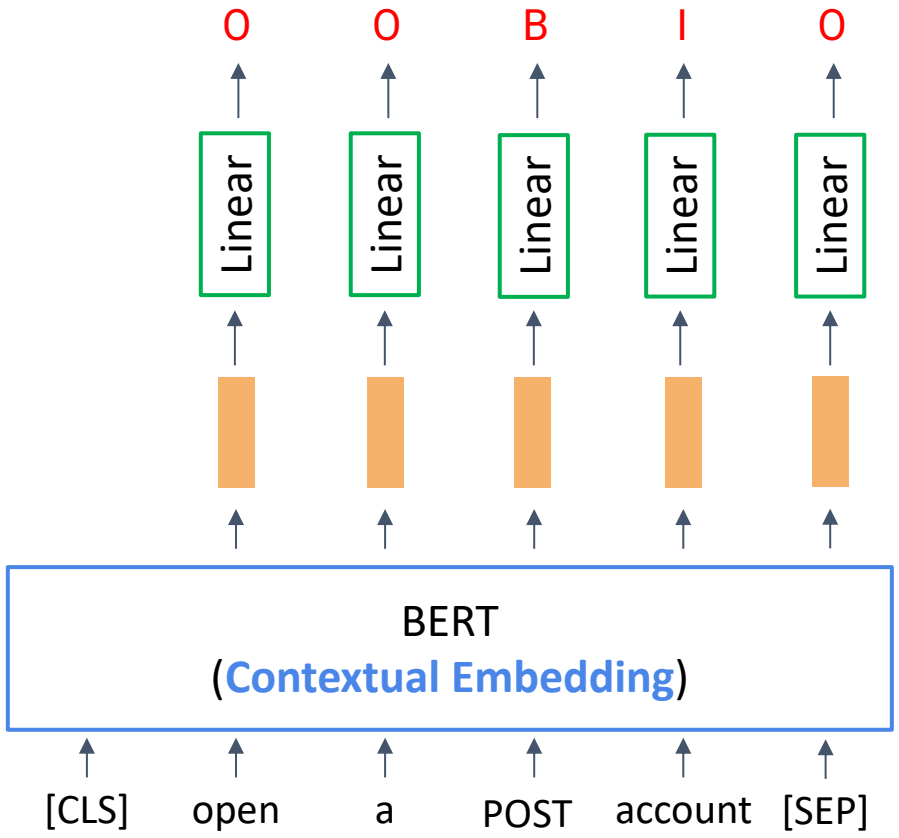
Question: To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?

Context: Architecturally, the school has a Catholic character. Atop the Main Building\'s gold dome is a golden statue of the Virgin Mary. Immediately in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend "Venite Ad Me Omnes". Next to the Main Building is the Basilica of the Sacred Heart. Immediately behind the basilica is the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858. At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome), is a simple, modern stone statue of Mary.

Answer: { "text": ["Saint Bernadette Soubirous"], "answer_start": [515] }

https://huggingface.co/docs/transformers/en/tasks/question_answering

Fine-tuning BERT (Named-entity Recognition)



RoBERTa: A Robustly Optimized BERT Pretraining Approach

- 和 BERT 的結構完全一樣
- 預訓練移除 Next Sentence Prediction (NSP)，且使用更多的資料進行預訓練
- Dynamic masking (BERT 全部的資料集只做一次隨機 masking)
- 更好的超參數 (batch size, learning_rate, num_epochs)
- 實務上確實效能通常比 BERT 還要好

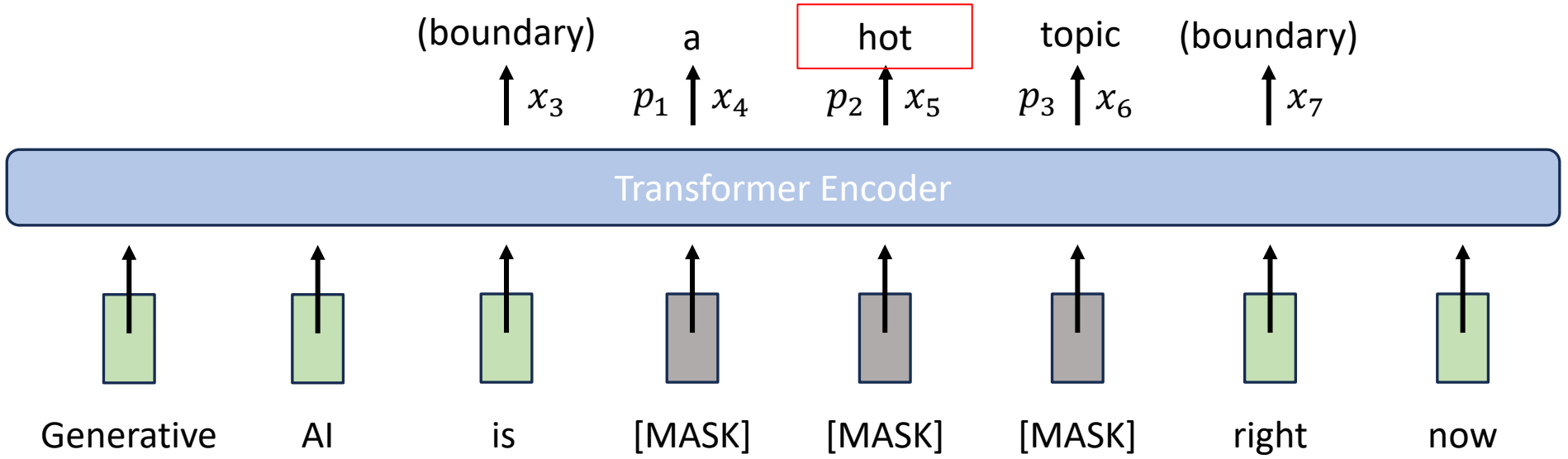
Rejected by ICLR 2020. https://openreview.net/forum?id=SyxS0T4tvS

SpanBERT

https://arxiv.org/pdf/1907.10529

基於MLM的新目標函數: Span Boundary Objective (SBO)

➡️ 一次 mask 一段 span，每次的目標為還原 span 中的其中一個字



$$\mathcal{L}(\text{hot}) = \mathcal{L}_{MLM}(\text{hot}) + \mathcal{L}_{SBO}(\text{hot}) = -logP(\text{hot}|\, x_5) - logP(\text{hot}|\, x_3, x_7, p_2)$$

Extensions of BERT (Summary)

Model	MLM task	NSP task	Release
BERT	Static masking	Sentence relationship	2018/10
RoBERTa	Dynamic masking	Remove NSP task	2019/07
SpanBERT	Span masking	Remove NSP task	2019/07

Announcements

Week 8: Mid-term Exam

- 13: 10 – 14:00 Project Introduction
- 14:10 – 16:00 Mid-term Exam
- 題目：純手寫、簡答題

Project checkpoints (暫定)

- Week 9: 確定各組的題目
- Week 11: 進度報告 PPT (5 pages)
- Week 13: 進度報告 PPT (5+5 pages), Presentations (selected teams)
- Week 15 – Week 16: Final presentations for all teams (maybe poster)

Thank you!

Instructor: 林英嘉

✉ yjlin@cgu.edu.tw

TA: 吳宣毅

✉ m1161007@cgu.edu.tw