2

Why are (L)LMs powerful?

- Size
- Attention mechanism
- Layers
- Etc.

Transfer Learning

Transfer Learning

What:

Store ‘language knowledge’ and ‘task knowledge’ in the parameters of a model.

Why:

Learn new tasks faster & better.

2

Classical Algorithms (e.g. Regression, SVM)

No prior
‘language knowledge’

No knowledge of semantic similarities between words like "*attack*", "*war*" and ”*tree*".

No prior
‘task knowledge’

No knowledge of tasks like "*Classify this text into ‘activist’ or ‘conservative’ rhetoric*".

2

Word Embeddings
(e.g. Word2Vec)

Provide '**language knowledge**':

- Represents "*attack*", "*war*" in similar static vectors

Attack ≈	0.1	0.8	...	-0.5	0.1
War ≈	0.2	0.8	...	-0.4	0.3
Tree ≈	0.7	-0.4	...	0.1	-0.7

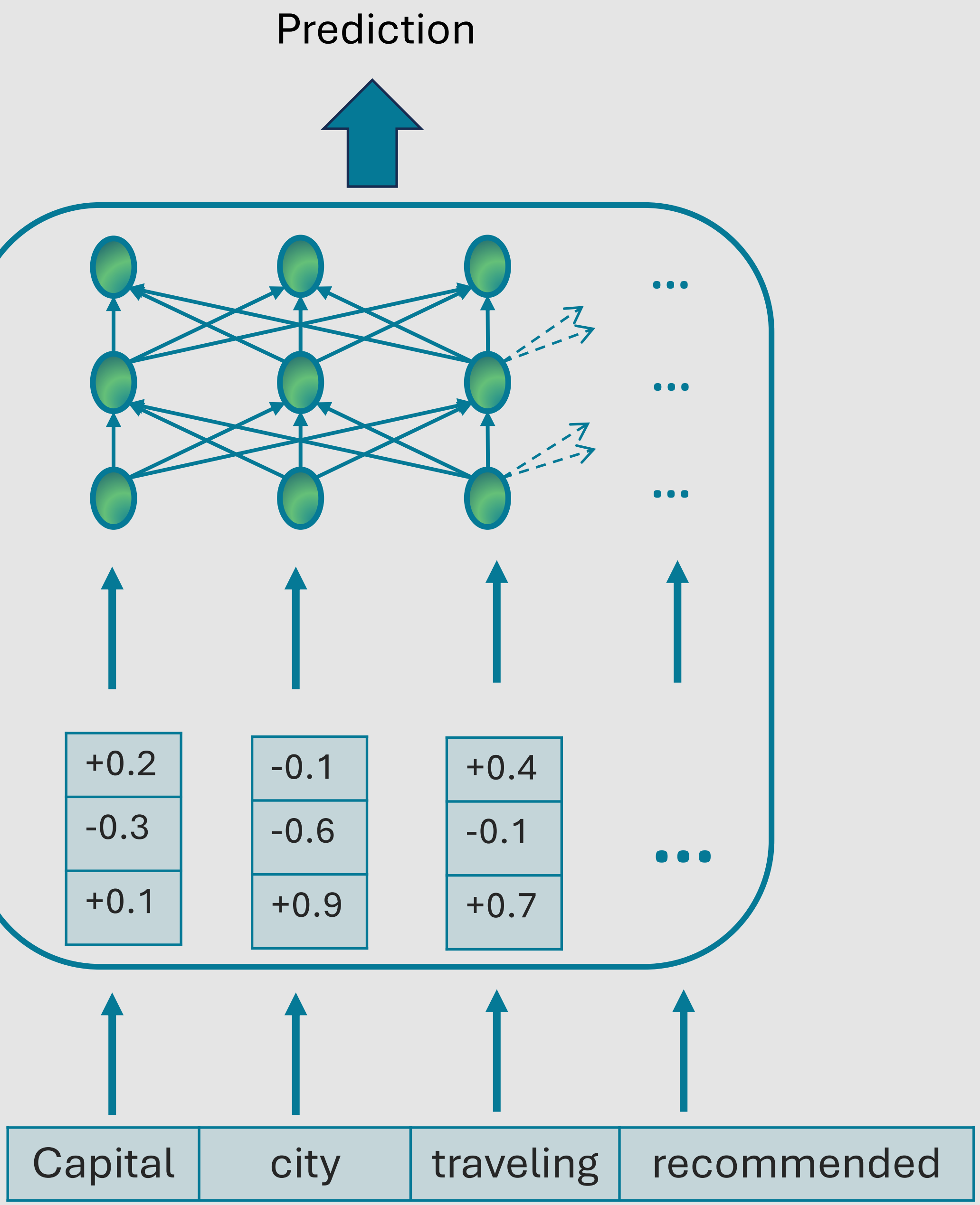
2

Transformers
(e.g. BERT-base)

Prior **'language knowledge**’:

- Represents "*attack*", "*war*" in similar vectors
- Represents "*capital*" differently in context of "*city*" or “*investment*” or “*crime*”

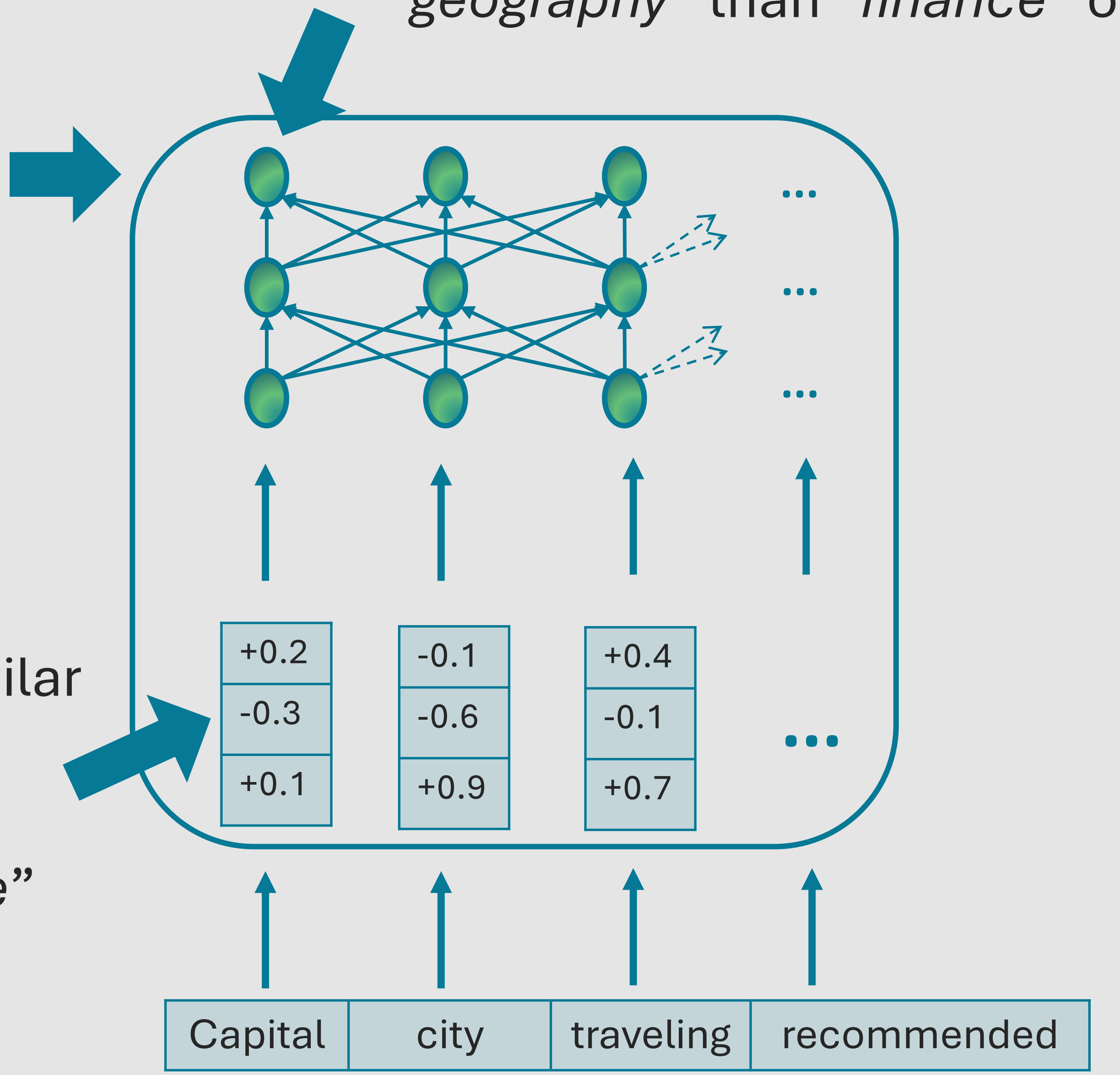
2


Prediction
...
...
...
+0.2
-0.3
+0.1
-0.1
-0.6
+0.9
+0.4
-0.1
+0.7
...
Capital
city
traveling
recommended

2

Word vector of “*capital*” now closer to “*geography*” than “*finance*” or “*crime*”

Last layer: Contextualised representation of input



Word vector of “*capital*” with similar distance to “*geography*” / “*finance*” / “*crime*”

2

Transformers
(e.g. BERT-base)

Prior '**language knowledge**’:

- Represents "*attack*", "*war*" in similar vectors
- Represents "*capital*" differently in context of "*city*" or “*investment*” or “*crime*”

Learned through **simple, self-supervised task**:

- Masked Language Modelling

2

How BERT acquires language knowledge

ORIGINAL TEXT

“Capital punishment, also known as the death penalty and formerly called judicial homicide, is ...”

MASKED TEXT

“Capital **[MASK]**, also known as the death **[MASK]** and formerly called **[MASK]** homicide, is ...”

The algorithm learns to predict the correct word behind the [MASK] token.

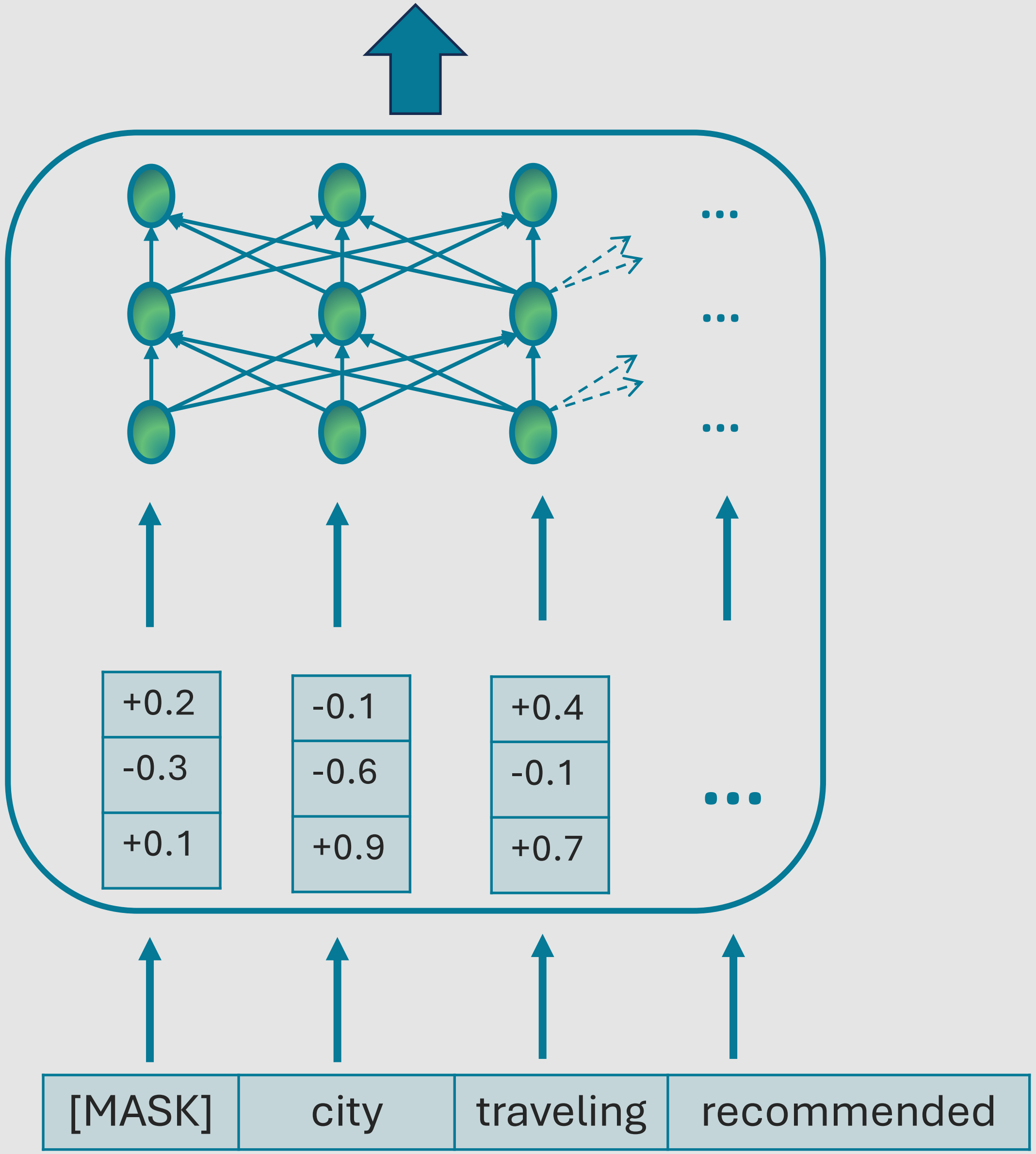
This creates general '**language knowledge**'

2

BERT-base

After millions of training iterations:

Many parameters (vectors), which are very good at predicting hidden words.



2

Disadvantages of BERT-base

‘Task knowledge’ from MLM is not useful:

- BERT-base only knows how to predict hidden words (MLM task)
- We are actually interested in other tasks like classification, summarization etc.
- BERT-base needs to learn new, useful tasks from scratch

2

Reusing more ‘prior knowledge’

Universal tasks

- Natural Language Inference (BERT-NLI)
- Next-word-prediction (GPT)

2

Natural Language Inference (NLI)

$\underbrace{\text{The EU betrayed its partners during the negotiations on Sunday.}}_{\textbf{context-sentence}}$

$\underbrace{\text{The EU is trustworthy.}}_{\textbf{hypothesis-sentence}}$

NLI Task

Is the **hypothesis** true/false/neutral, given the **context-sentence**?

→ NLI is the task of determining whether a given statement (hypothesis) can logically be inferred from another statement (context-sentence).

2

Natural Language Inference (NLI)

INPUT

CONTEXT-HYPOTHESIS PAIR

The EU betrayed its partners during the negotiations on Sunday. [SEP] The EU is trustworthy.

context-sentence **hypothesis-sentence**

OUTPUT

PROBABILITY OF 3 UNIVERSAL CLASSES

True: 0.01 **False: 0.83** Neutral: 0.26

2

Universal Task & Label Verbalisation

Example task:

Identifying texts that indicate that the economy is performing well / badly

Task reformulated for NLI:

NLI-input	*NLI-Output*
{context-sentence from news} [SEP] {hypothesis-sentence verbalising label}	Most "True" label

"The Rubel plummeted amidst a surge of investors withdrawing from Russia [SEP] The economy is performing badly"	**0,61 True** 0,13 False 0,26 Neutral
"The Rubel plummeted amidst a surge of investors withdrawing from Russia [SEP] The economy is performing well"	0,19 True 0,38 False 0,43 Neutral

2

Universal Task & Label Verbalisation

NLI-input	*NLI-Output*
{context-sentence from news} [SEP] {hypothesis-sentence verbalising label}	Most "True" label

"The politicians were bribed by lobbyists. [SEP] It is about corruption."	**0,61**
"The politicians were bribed by lobbyists. [SEP] It is about peace."	0,01
"The politicians were bribed by lobbyists. [SEP] It is about free market."	0,06
"The politicians were bribed by lobbyists. [SEP] It is about equality."	0,04
...	...

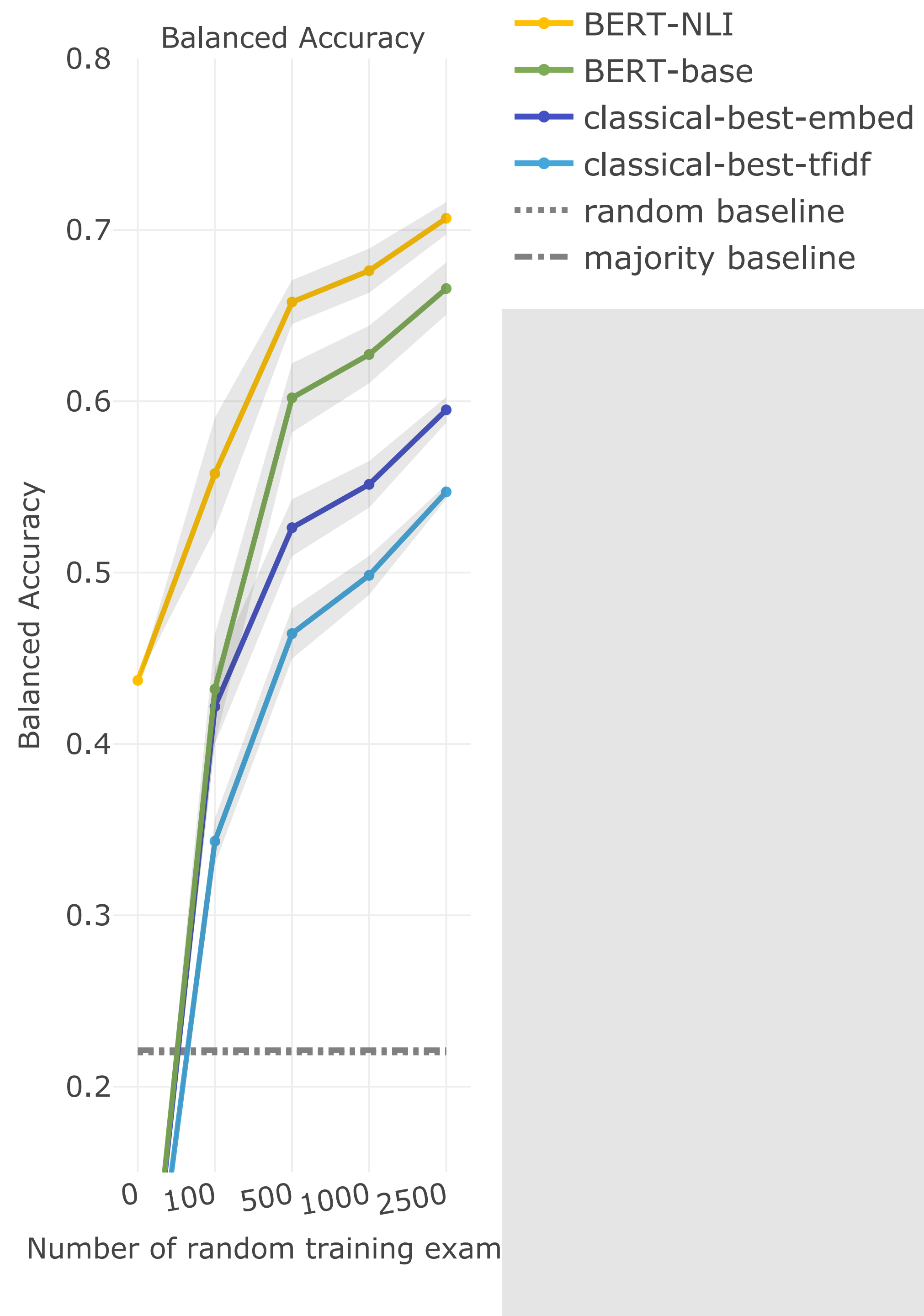
NLI: Data-Rich Task

NLI is a data rich task

- Many NLI datasets with over 1 million annotated sentence pairs from different domains exist.

- Examples: *SNLI (570k examples, Bowman et al. 2015), MultiNLI (433k, Williams et al. 2018), ANLI (162k, Nie et al. 2020)*

→ **Helps address the issue of data scarcity**



Average performance across eight tasks vs. training data size

[Laurer et. al 2023]

Limitations of NLI

- Usefulness decreases with training data size. If there is enough data to learn the new task (> 1000 texts), BERT-base is better.
- BERT-NLI can only do classification tasks.
- No summarization, translation, information extraction ...

A more universal task

Next-token-prediction

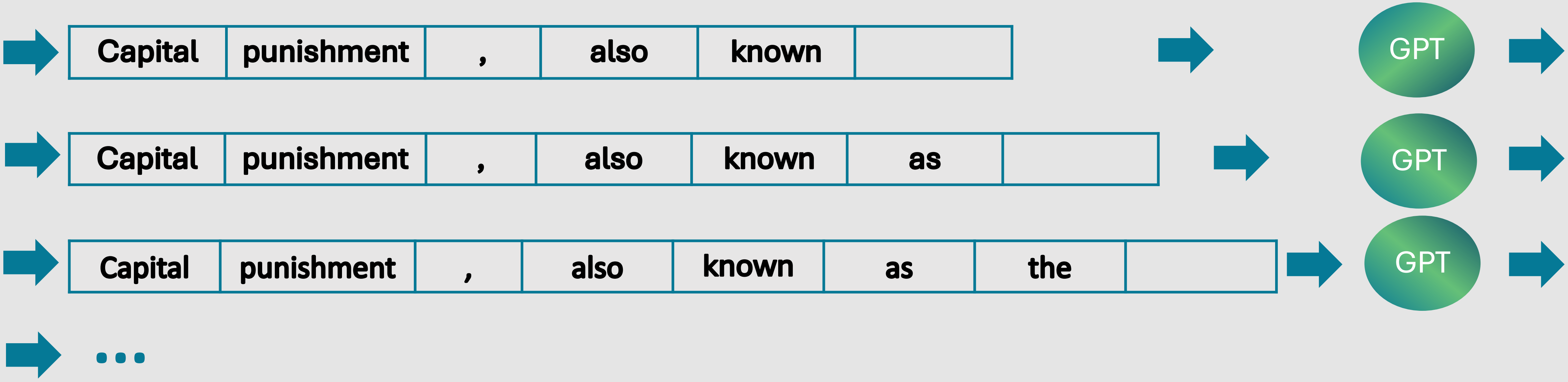
- Main pre-training task of GPT models
- Is a self-supervised task (no manual annotations)

2

Next-token-prediction task

Original text:

“Capital punishment, also known as the death penalty and formerly called judicial homicide, is …”



Universal task

Sentiment classification

[long text ...]	Is	this	text	positive	or
negative	?				

GPT

[long text ...]	Is	this	text	positive	or
negative	?	**positive**			

2

Universal task

Information extraction

[long text ...]	Extract	all	countries	from	the
text	:				

GPT

[long text ...]	Extract	all	countries	from	the
text	:	**Germany**			

GPT

[long text ...]	Extract	all	countries	from	the
text	:	**Germany**	,		

GPT

...

2

Universal task

Summarisation

[long text ...]	A	summary	of	the	preceding
text	:				

GPT

[long text ...]	A	summary	of	the	preceding
text	:	**The**			

GPT

[long text ...]	A	summary	of	the	preceding
text	:	**The**	**main**		

GPT

...

Reflect and Q&A

- Q1: What is the difference between word embeddings and BERT?
- Q2: What are universal tasks and why are they useful?
- Q3: In your own words, try to define transfer learning.

Write your responses on a piece of paper / notebook.
Ask any questions about the slides in the chat.