自督導式學習的神奇能力

李宏毅

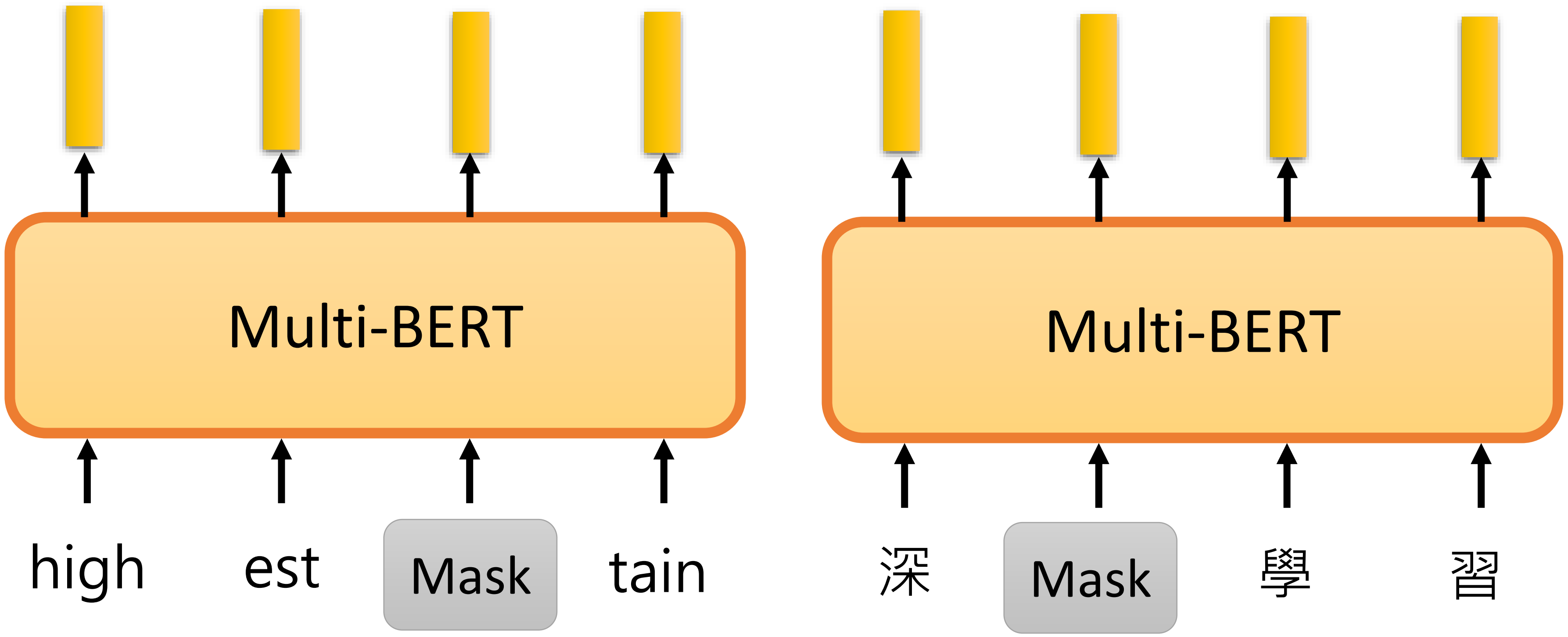
Outline

Story 1: Cross-lingual

Story 2: Cross-discipline

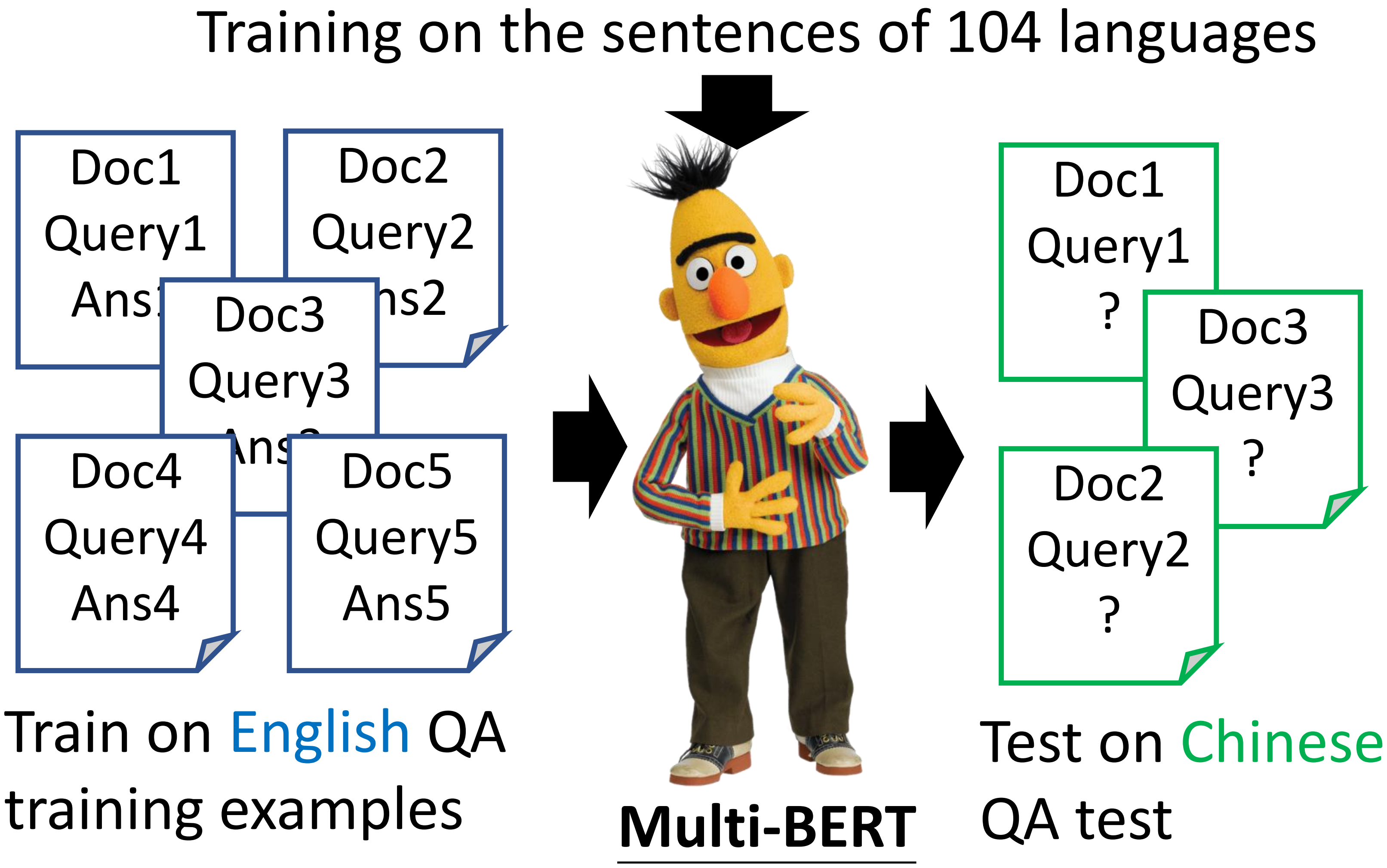
Story 3: Pre-training without Human Languages

Multi-lingual BERT



Training a BERT model by many different languages.

Zero-shot Reading Comprehension



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%

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

So many evidences ……

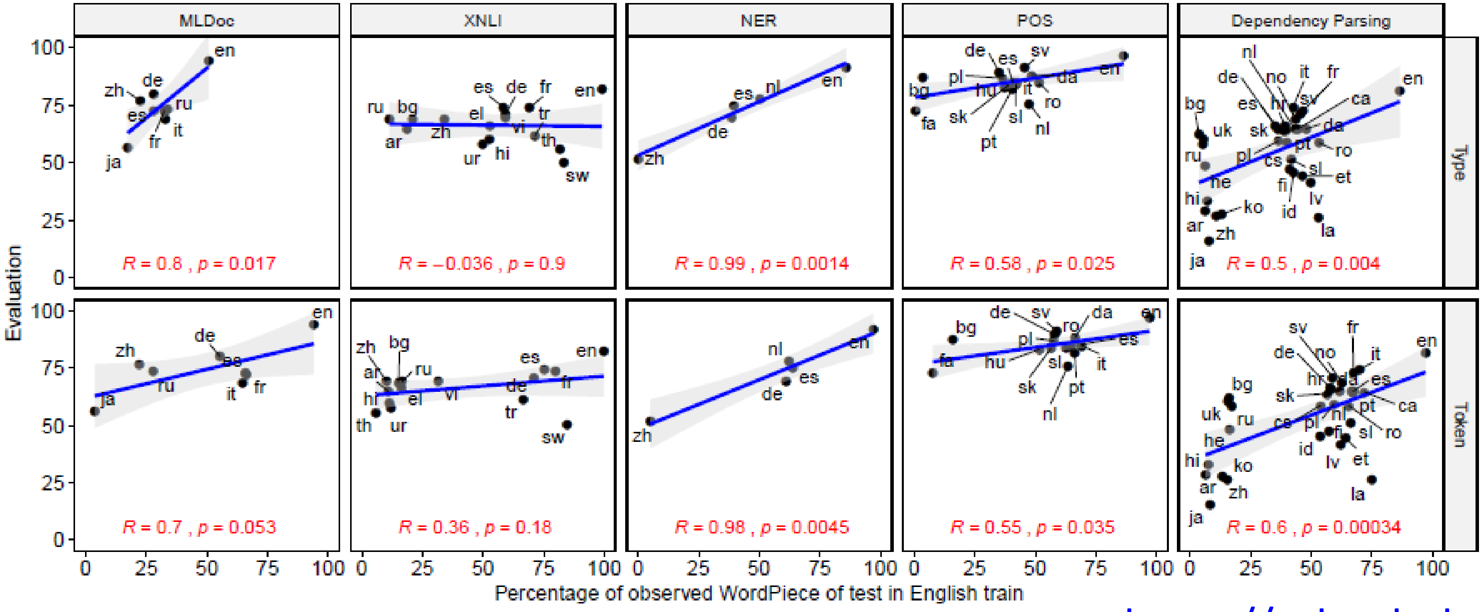
Fine-tuning \ Eval	EN	DE	NL	ES
EN	**90.70**	69.74	77.36	73.59
DE	73.83	**82.00**	76.25	70.03
NL	65.46	65.68	**89.86**	72.10
ES	65.38	59.40	64.39	**87.18**

Table 1: NER F1 results on the CoNLL data.

Fine-tuning \ Eval	EN	DE	ES	IT
EN	**96.82**	89.40	85.91	91.60
DE	83.99	**93.99**	86.32	88.39
ES	81.64	88.87	**96.71**	93.71
IT	86.79	87.82	91.28	**98.11**

Table 2: POS accuracy on a subset of UD languages.

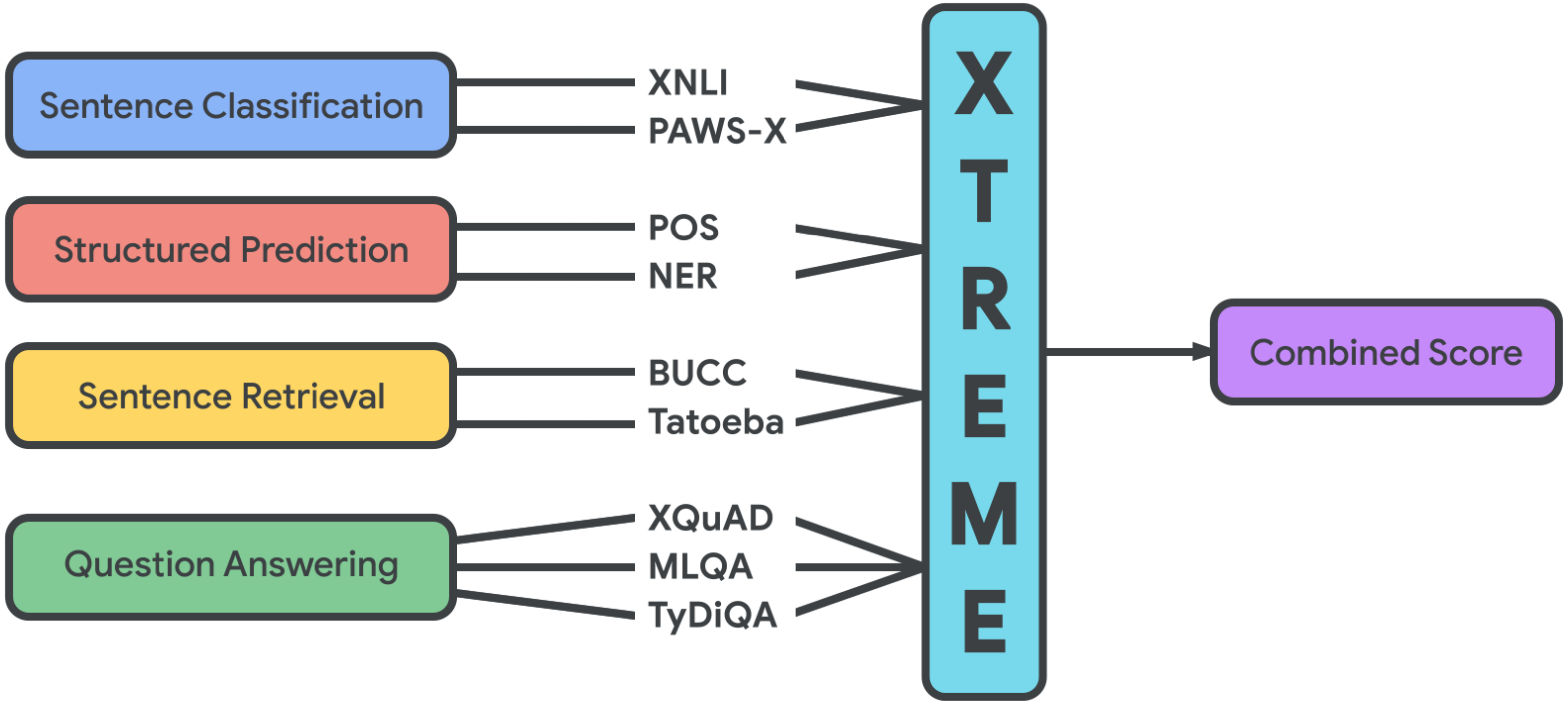
https://aclanthology.org/P19-1493/



https://aclanthology.org/D19-1077/

Cross-lingual TRansfer Evaluation of Multilingual Encoders (XTREME) benchmark

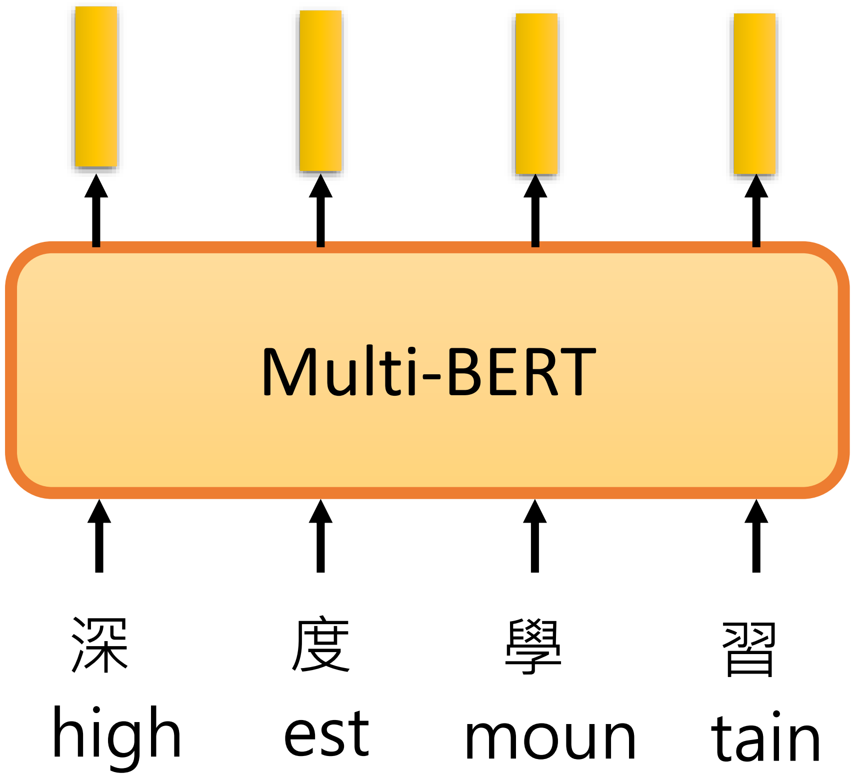
https://sites.research.google/xtreme

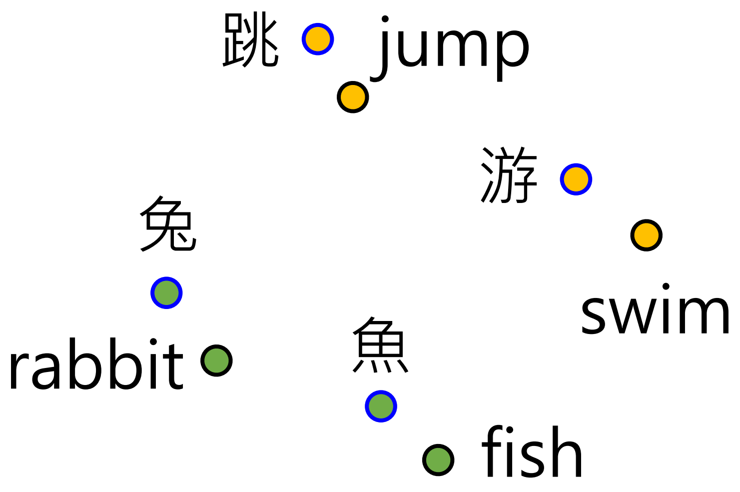


40 languages for 9 tasks

Train on English, and test on the rest

Cross-lingual Alignment?





How alignment happens?

- Typical answer

Different languages share some common tokens.

How do you explain Chinese v.s. English?

Code Switching

… DNA 的構造很像螺旋梯 …
(digits, punctuations)

Intermediate Language?

Language X shares tokens with Chinese and English.

How alignment happens?

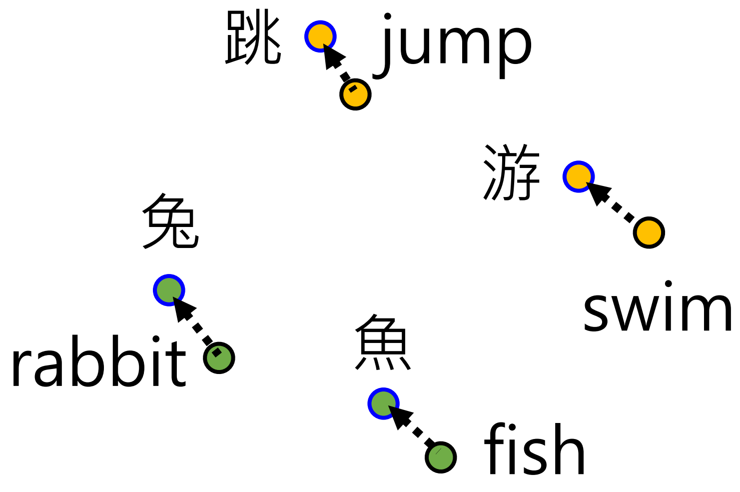
https://openreview.net/forum?id=HJeT3yrtDr

			XNLI		NER
B-BERT	Train	Test	Accuracy	Wordpiece Contribution	Span F1-Score
en-es	en	es	72.3	1.4	61.9 (±0.8)
enfake-es	enfake		**70.9**		**62.6 (±1.6)**
en-hi	en	hi	60.1	0.5	61.6 (±0.7)
enfake-hi	enfake		**59.6**		**62.9 (±0.7)**
en-ru	en	ru	66.4	0.7	57.1* (±0.9)
enfake-ru	enfake		**65.7**		**54.2 (±0.7)**
en-enfake	enfake	enfake	78.0	0.5	78.9* (±0.7)
en-enfake	enfake	en	**77.5**		76.6(±0.8)

English:	the	cat	is	a	good	cat
Fake-English:	甲	乙	天	地	人	乙

https://arxiv.org/abs/2010.10041

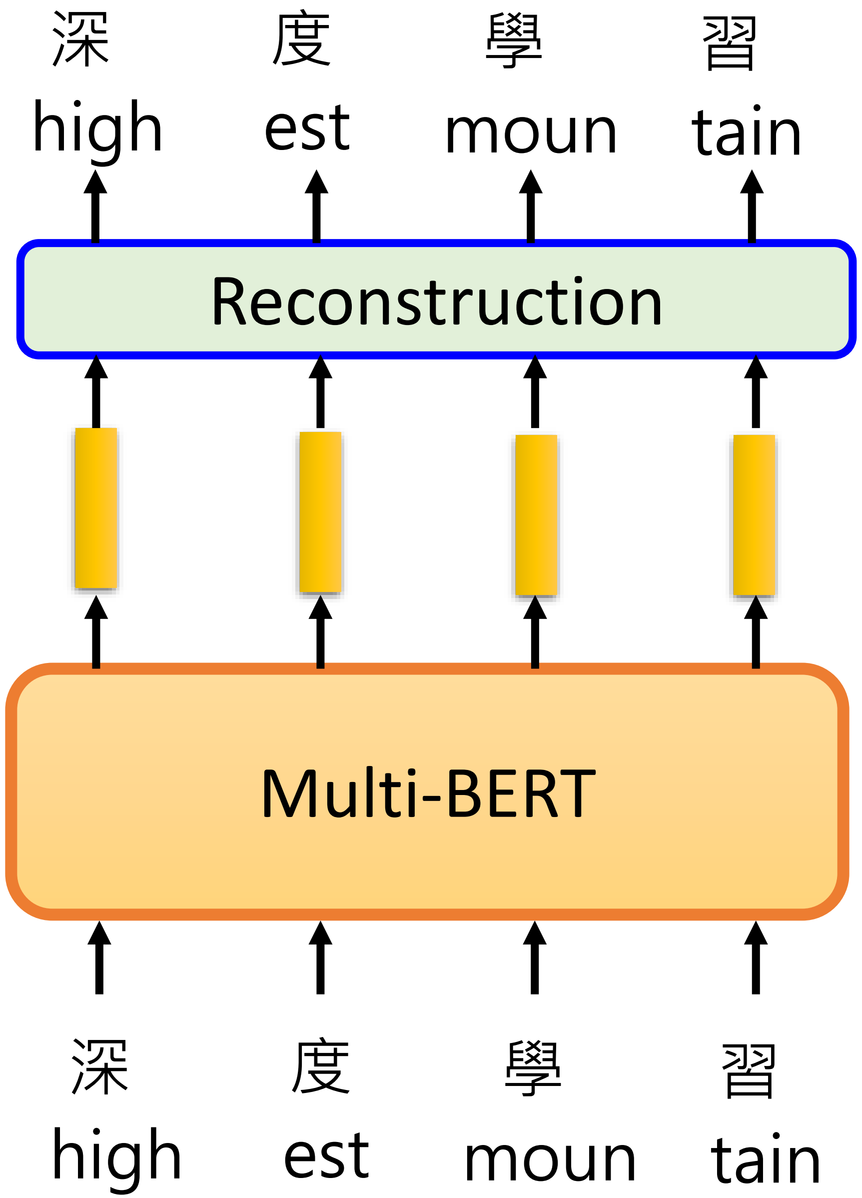
Weird???



If the embedding is language independent …

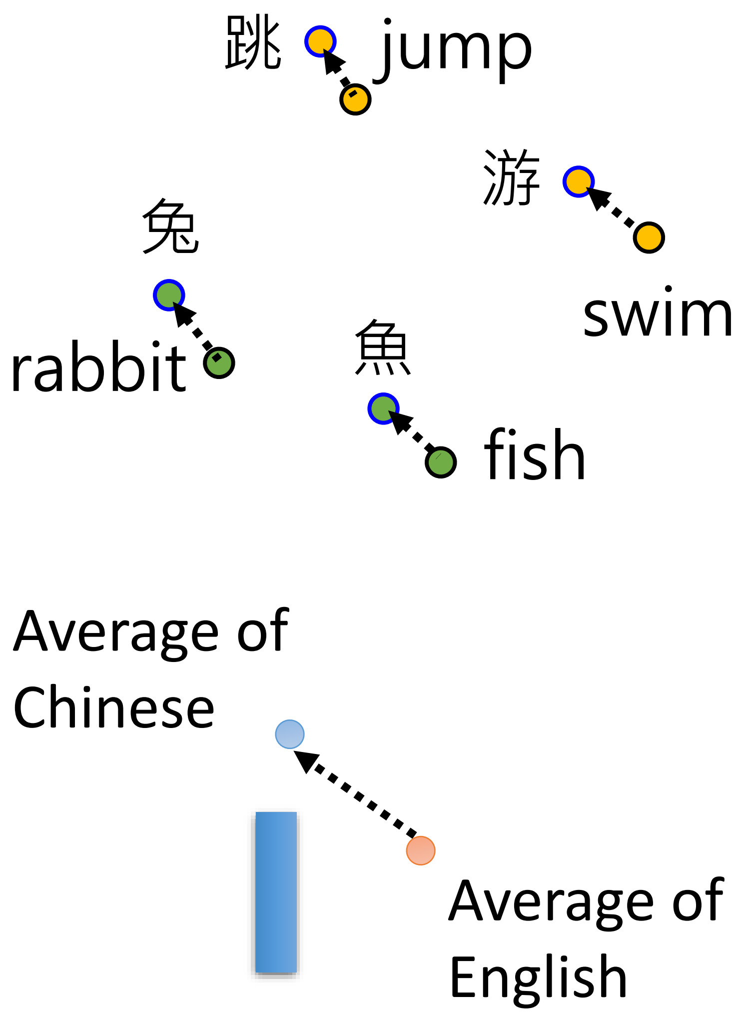
How to correctly reconstruct?

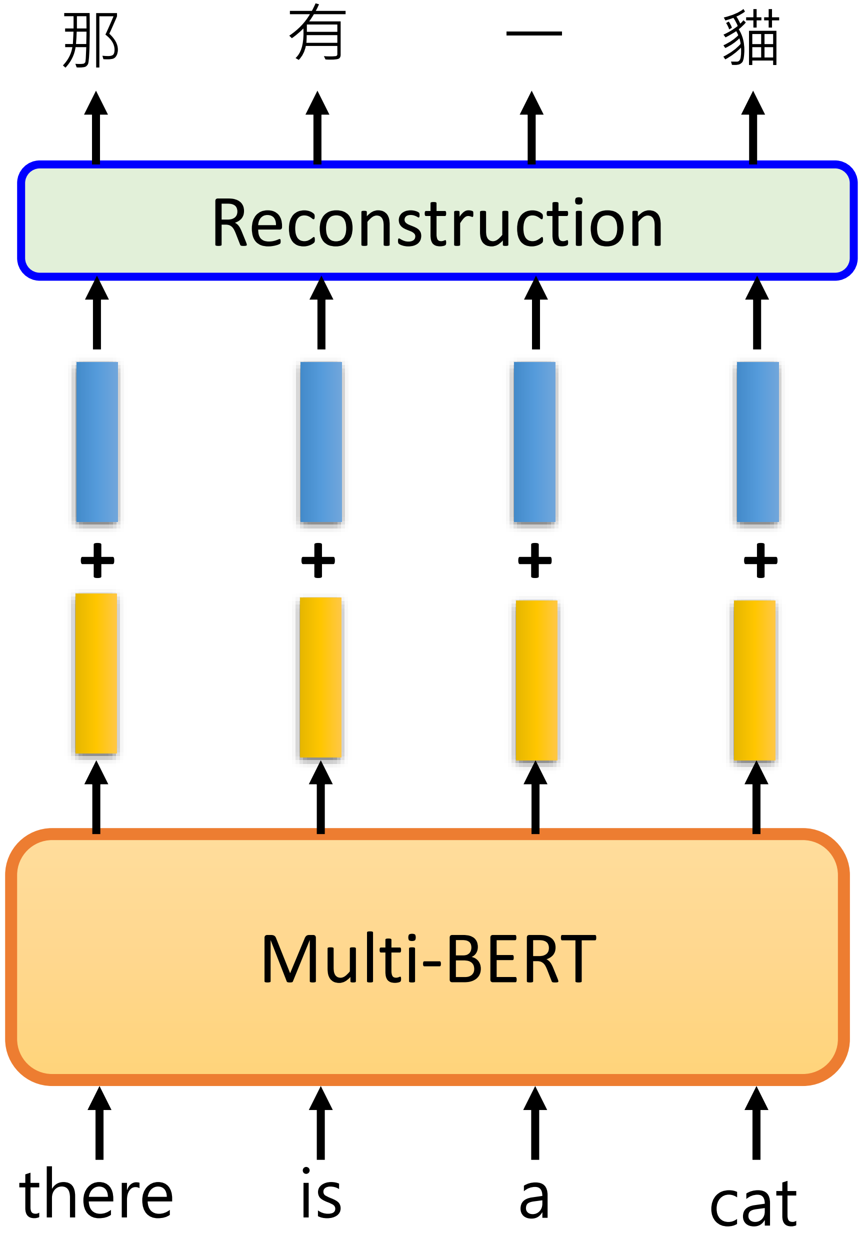
There must be language information.



Where is Language?

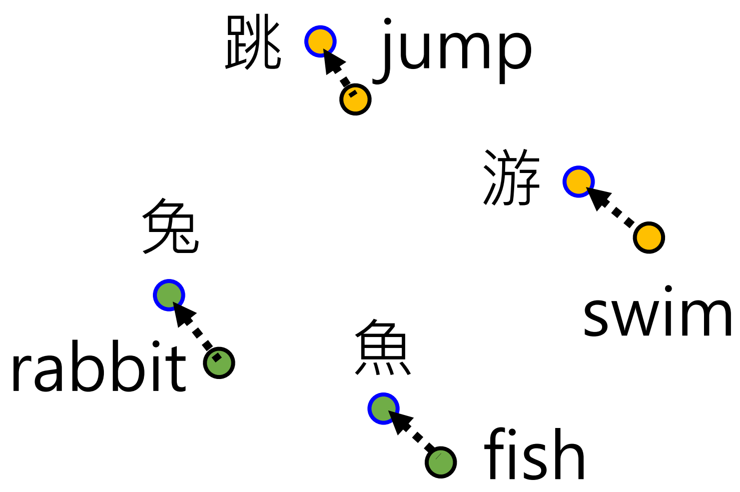
This work is done by 劉記良、許宗頫、莊永松

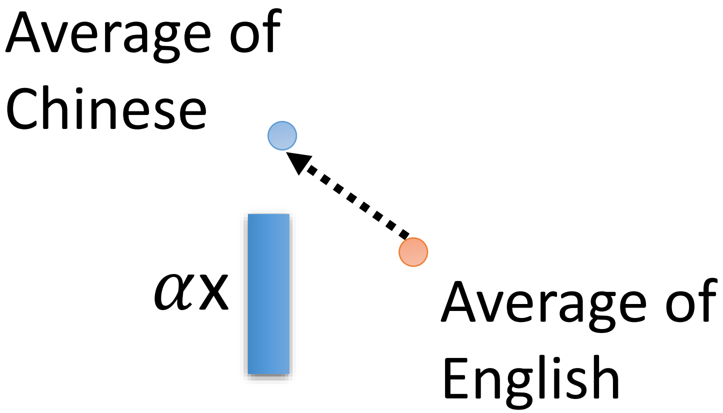




This work is done by 劉記良、許宗嫄、莊永松
https://arxiv.org/abs/2010.10041

If this is true …





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 ☺

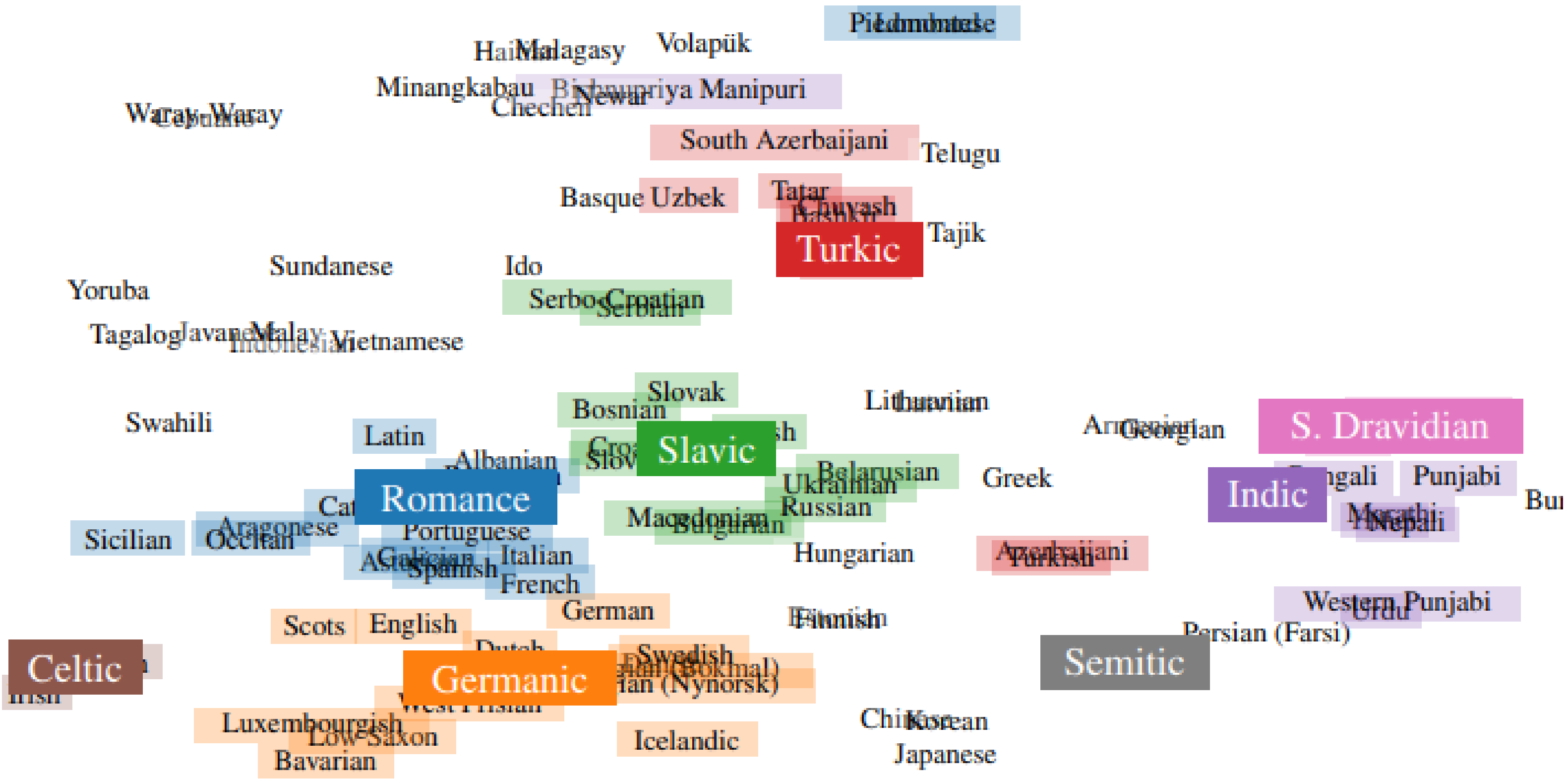
This work is done by 劉記良、許宗嫄、莊永松
https://arxiv.org/abs/2010.10041

Table 1: Unsupervised Token Translation quantitative results using the 10-th layer of BERT.

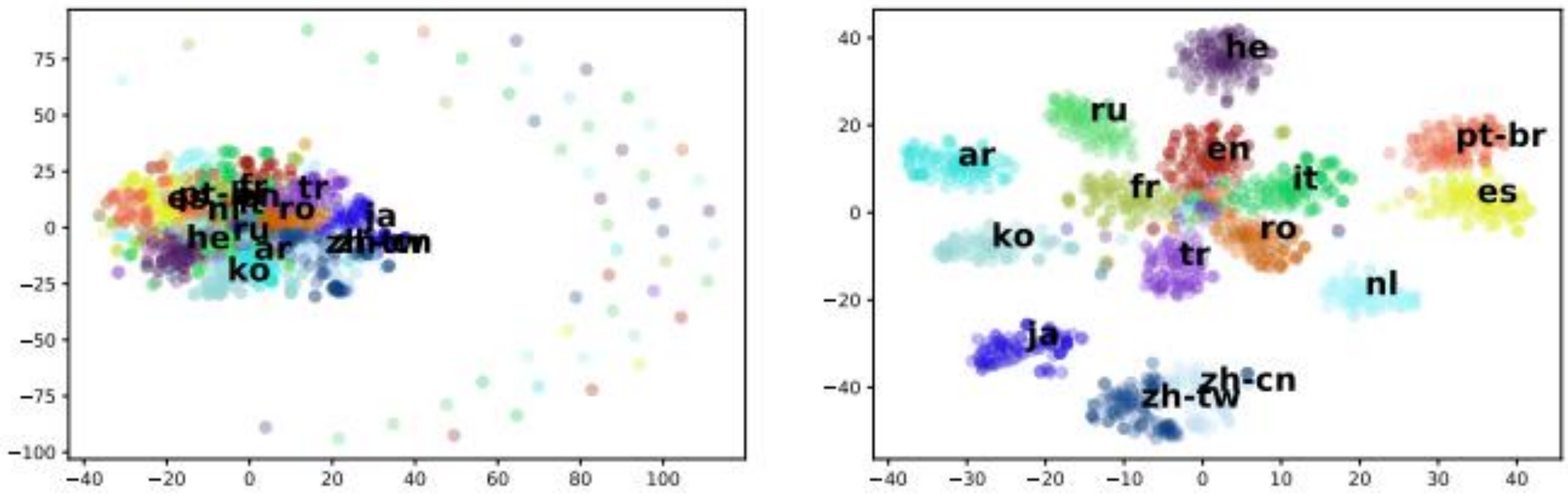
	en→de	en→fr	en→ur	en→sw	en→zh	en→el	de→en	fr→en	ur→en	sw→en	zh→en	el→en
BLEU-1 (α=1)	7.53	8.53	5.56	7.96	15.25	7.88	7.34	9.08	5.52	6.34	4.37	6.54
BLEU-1 (α=2)	8.03	10.24	6.31	7.23	21.51	14.91	7.48	8.52	6.23	7.48	5.38	6.65
BLEU-1 (α=3)	12.35	10.65	5.35	7.16	15.95	19.13	6.29	12.27	5.74	6.45	6.17	4.73
convert rate (α=1)	40.2	41.7	61.1	15.3	47.8	62.1	45.2	49.6	29.9	14.7	23.9	30.2
convert rate (α=2)	74.8	75.7	99.4	97.4	90.0	99.1	67.3	601	83.0	65.6	60.8	97.9
convert rate (α=3)	95.2	96.3	99.8	100	99.5	100	79.5	73.1	96.6	93.6	91.4	99.7

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 ☺

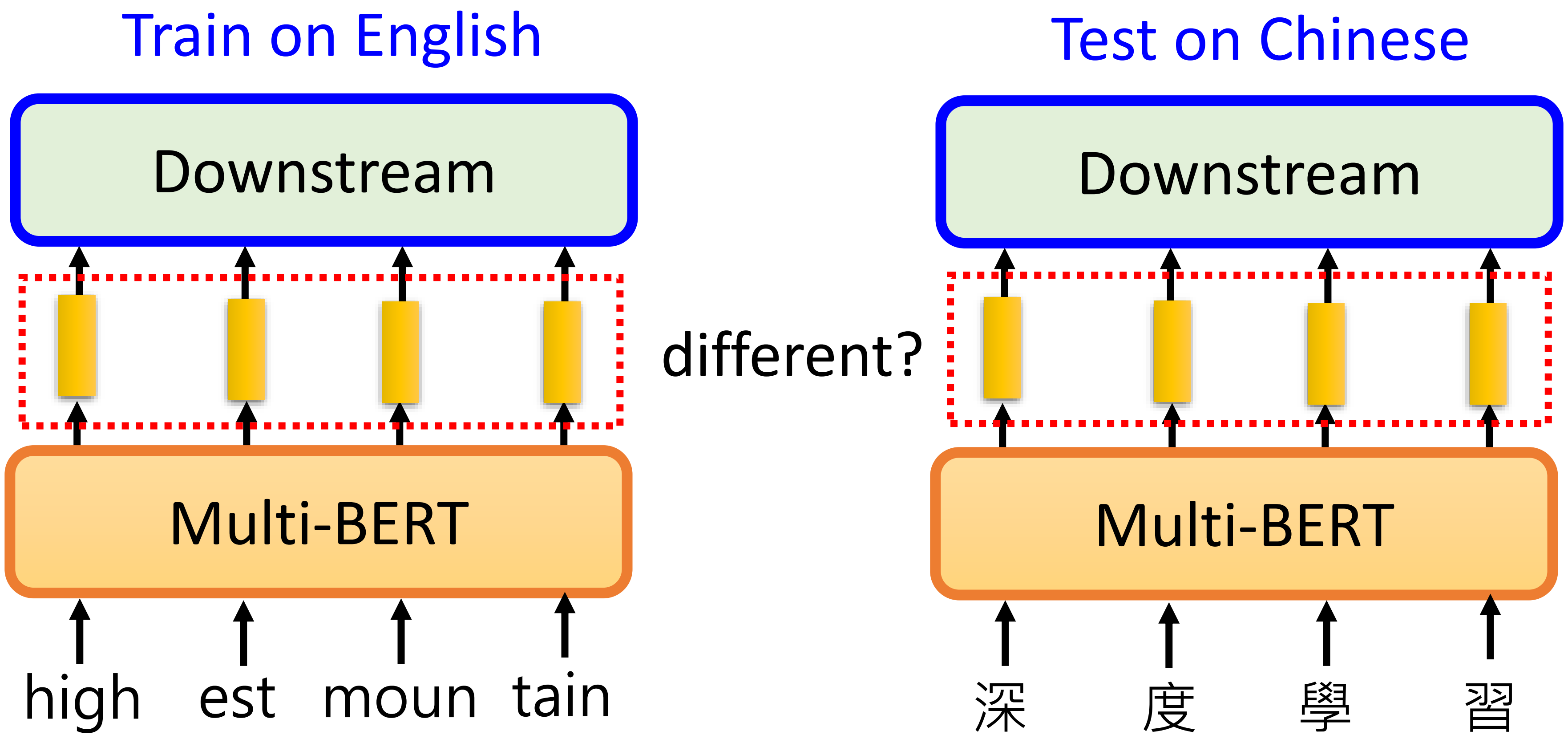


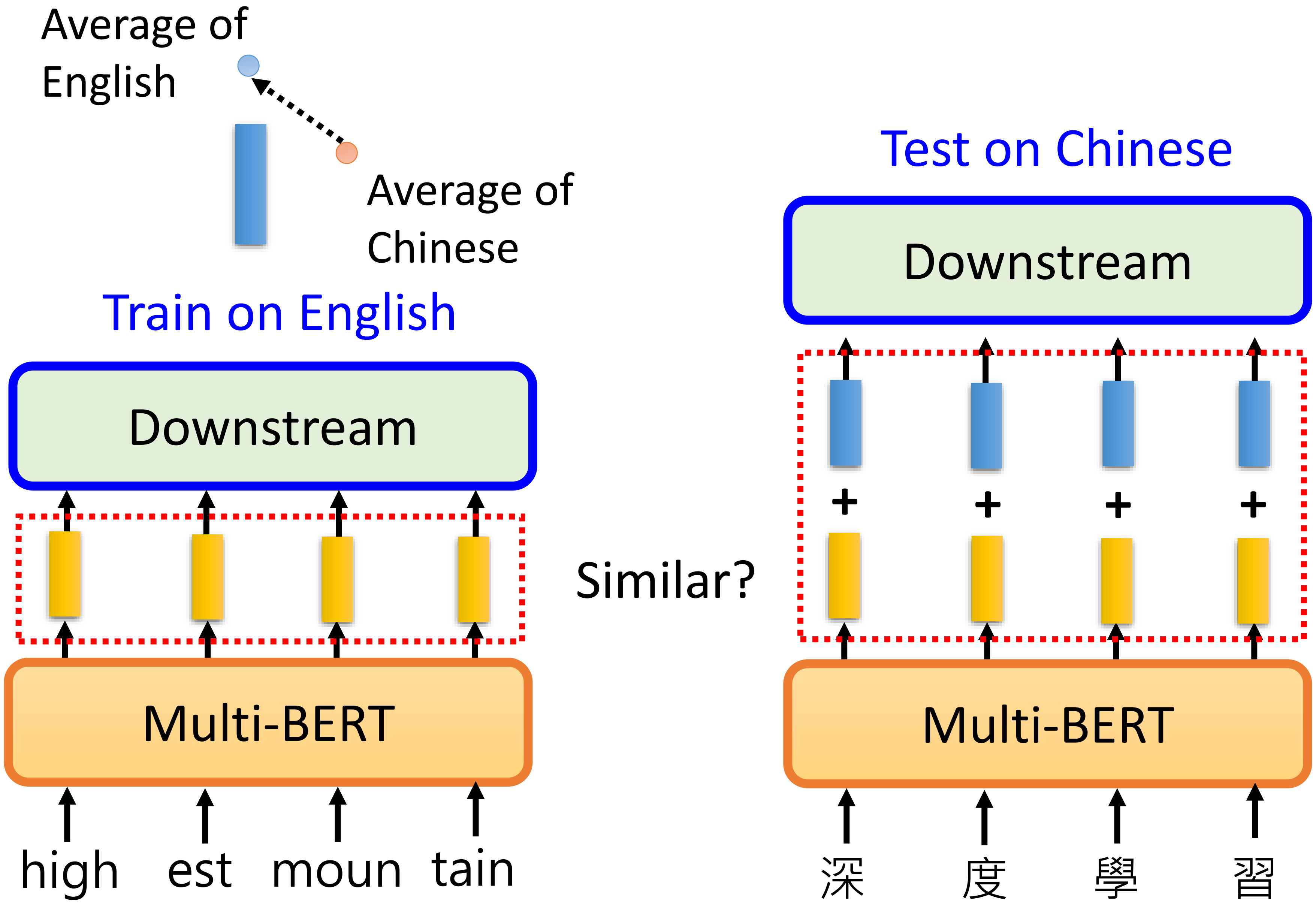
On the Language Neutrality of Pre-trained Multilingual Representations
https://arxiv.org/abs/2004.05160



It’s not Greek to mBERT: Inducing Word-Level Translations from Multilingual BERT
https://arxiv.org/abs/2010.08275

Zero-shot Cross-lingual Transfer




Average of
English
Average of
Chinese
Train on English
Downstream
Similar?
Multi-BERT
high
est
moun
tain
Test on Chinese
Downstream
+
+
+
+
Multi-BERT
深
度
學
習

Experimental Results

Table 4: POS tagging results

Method	ar	bg	de	el	es	fr	hi	ru	th	tr	ur	vi	zh	Average
Original	53.8	85.4	86.2	81.1	86.1	42.9	66.8	85.5	41.7	68.6	56.3	**53.8**	61.8	66.9
Zero-mean	**54.3**	86.1	**86.6**	**81.8**	86.6	43.7	68.1	**86.5**	41.6	**69.7**	56.6	53.4	62.5	67.5
MDS	54.2	**86.4**	86.5	81.5	**86.8**	**43.9**	**68.9**	86.4	**44.2**	69.4	**57.1**	52.4	**63.0**	**67.8**

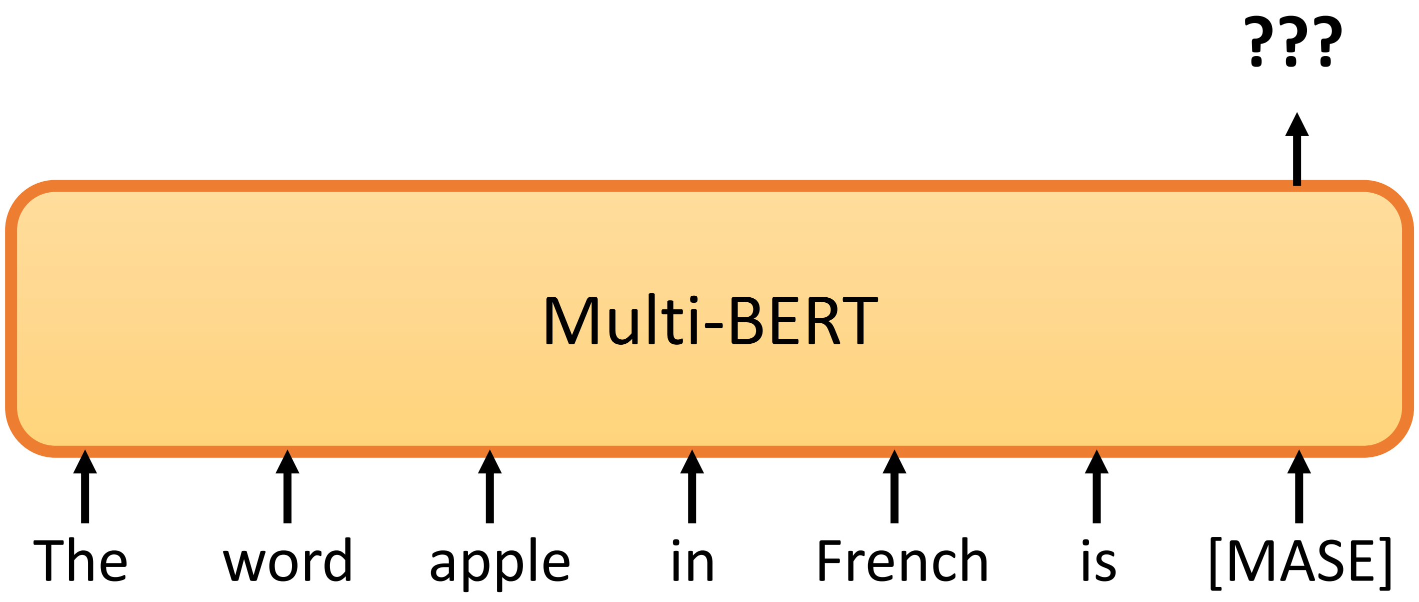
Table 5: Dependency parsing results. Numbers are Labeled Attachment Score(LAS).

Method	ar	bg	de	el	es	fr	hi	ru	th	tr	ur	vi	zh	Average
Original	28.2	70.7	**74.0**	**71.6**	72.1	74.8	35.3	69.0	30.8	32.9	28.3	**37.8**	**35.4**	50.8
Zero-mean	**28.2**	**71.0**	73.4	71.4	72.2	**75.7**	36.3	**69.3**	**32.5**	**34.6**	28.6	37.0	35.2	**51.2**
MDS	28.0	70.8	73.7	71.1	**72.2**	75.3	**36.5**	68.8	30.4	34.2	**29.0**	35.6	35.0	50.8

You just have to ask ……

It’s not Greek to mBERT: Inducing Word-Level Translations from Multilingual BERT
https://arxiv.org/abs/2010.08275

	@1	@10	@100
Baseline	0.036	0.244	0.575
Analogies	0.105	0.463	0.737
Template	**0.449**	**0.703**	**0.845**



Outline

Story 1: Cross-lingual

Story 2: Cross-discipline

Story 3: Pre-training without Human Languages

https://arxiv.org/abs/2103.07162

This work is done by 高瑋聰

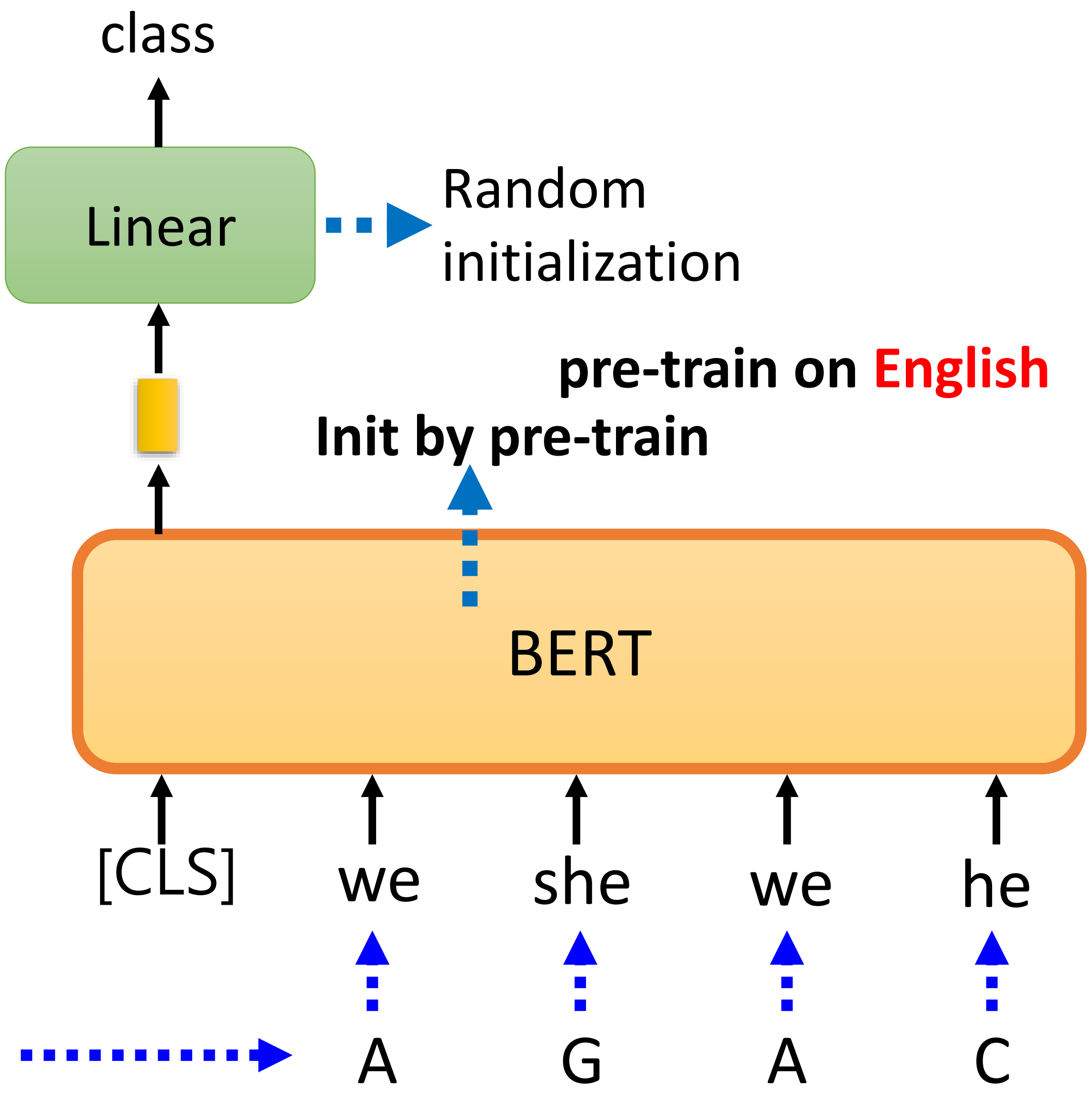
Why does BERT work?

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

class	DNA sequence
EI	CCAGCTGCATCACAGGAGGCCAGCGAGCAGGTCT
EI	AGACCCGCCGGGAGGCGGAGGACCTGCAGGGTG
IE	AACGTGGCCTCCTTGTGCCCTTCCCCACAGTGCCC
IE	CCACTCAGCCAGGCCCTTCTTCTCCTCCAGGTCCC
IE	CCTGATCTGGGTCTCCCCTCCCACCCTCAGGGAGC
IE	AGCCCTCAACCCTTCTGTCTCACCCTCCAGCCTAAA
IE	CCACTCAGCCAGGCCCTTCTTCTCCTCCAGGTCCC
N	CTGTGTTCACCACATCAAGCGCCGGGACATCGTGC
N	GTGTTACCGAGGGCATTTCTAACAGTCTTCTTACTA
N	TCTGAGCTCTGCATTTGTCTATTCTCCAGCTGACCC

Why does BERT work?

https://arxiv.org/abs/2103.07162
This work is done by 高瑋聰



A	we
T	you
C	he
G	she

DNA sequence

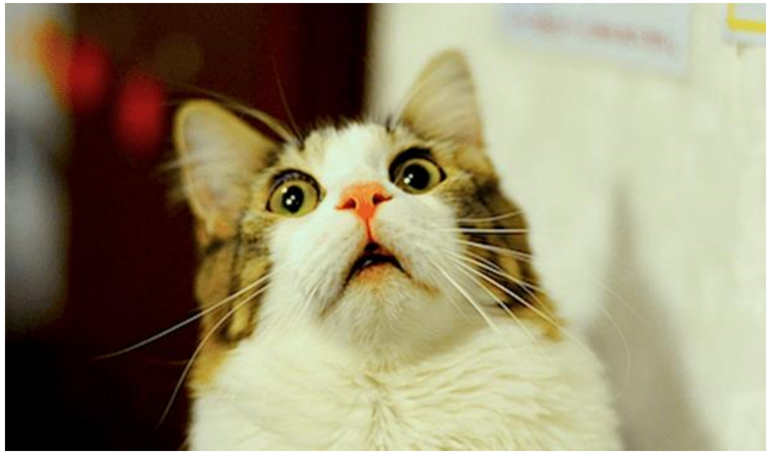
https://arxiv.org/abs/2103.07162

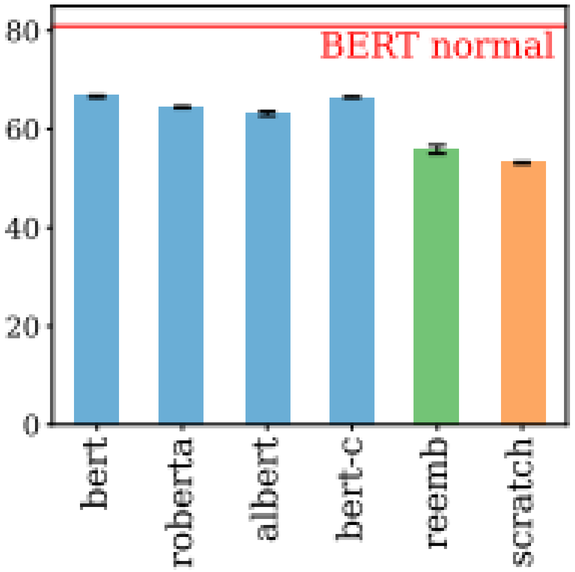
This work is done by 高瑋聰

Why does BERT work?

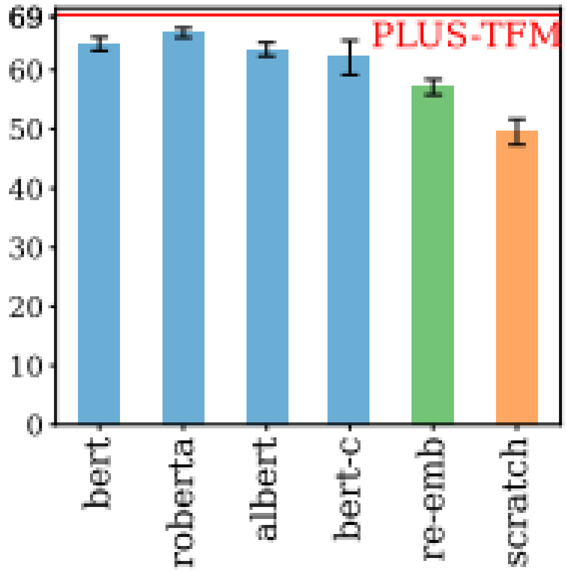
- Applying BERT to **protein, DNA, music classification**

	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

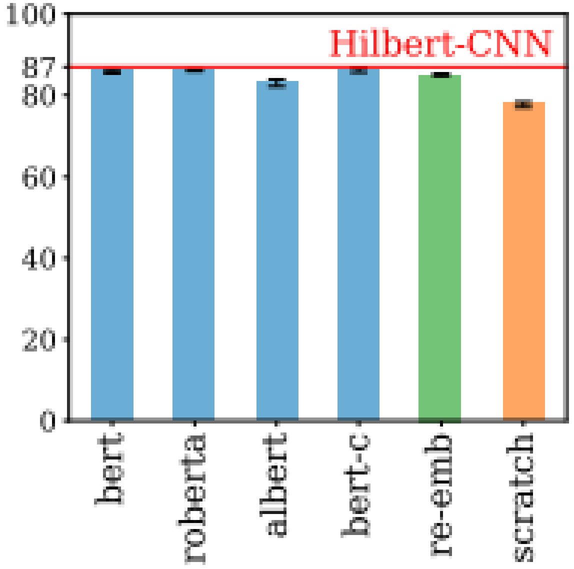




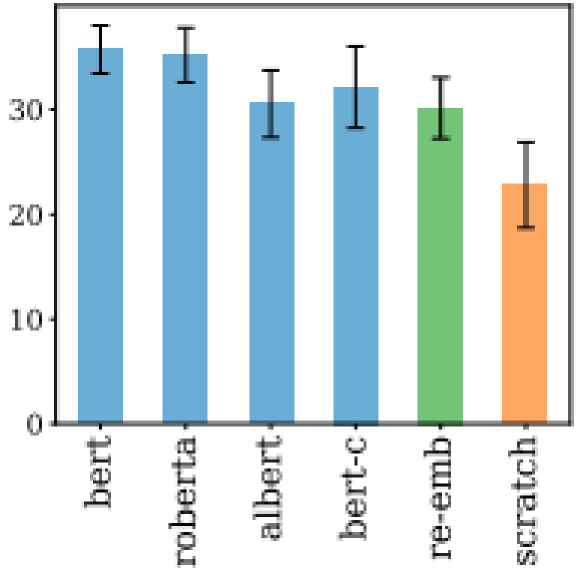
(a) Synthetic GLUE (8 tasks)



(b) Protein (3 tasks)



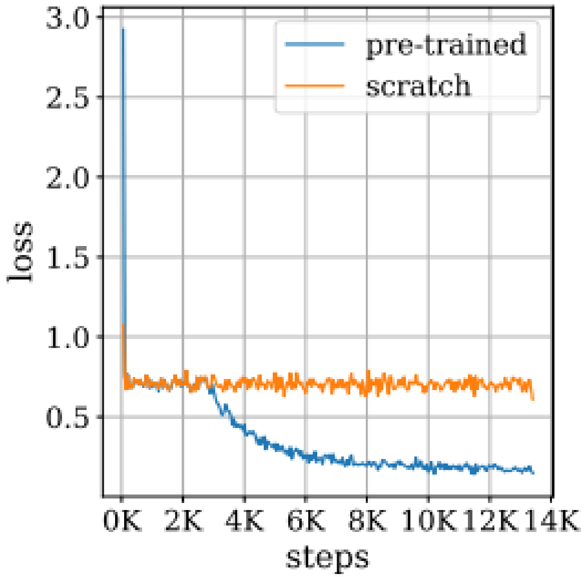
(c) DNA (4 tasks)



(d) music (1 task)

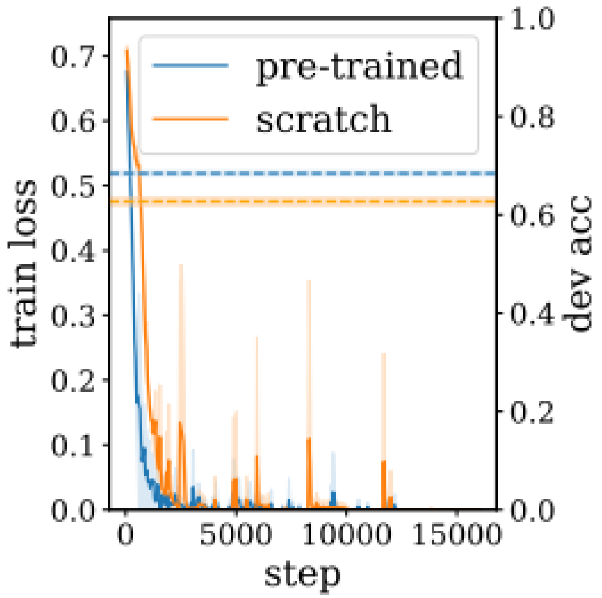
	Flu.	Stab.	Loc.
BERT - PLUS	**0.729**	**0.634**	**0.504**
BERT - random	0.598	0.545	0.362
PLUS - random	0.461	0.405	0.322
random - random	0.434	0.388	0.387

Optimization

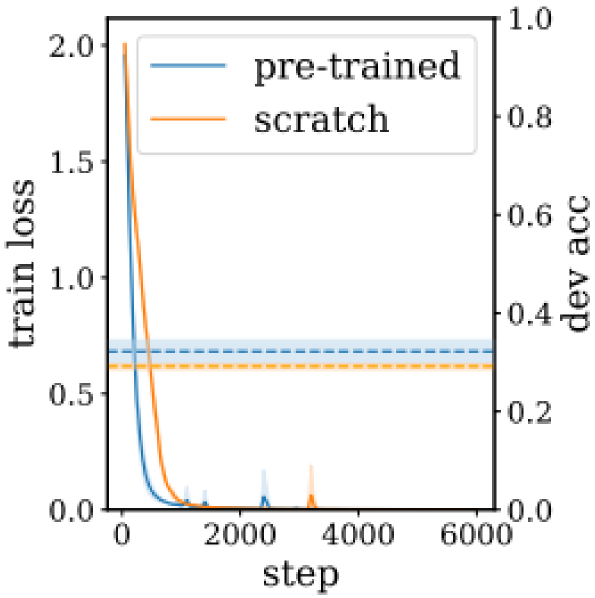


(b) fluorescence

Generalization

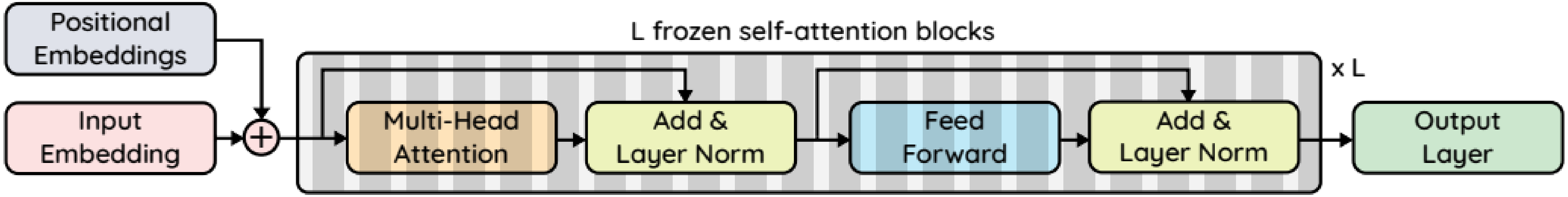


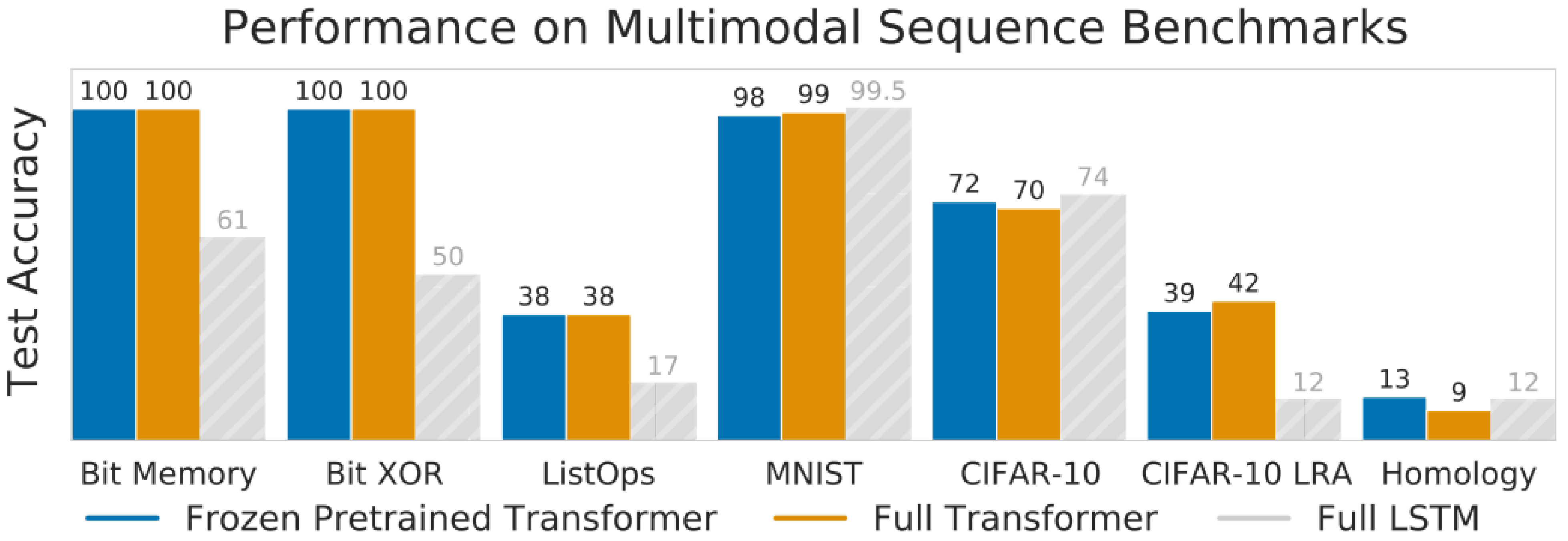
(a) H3K9ac

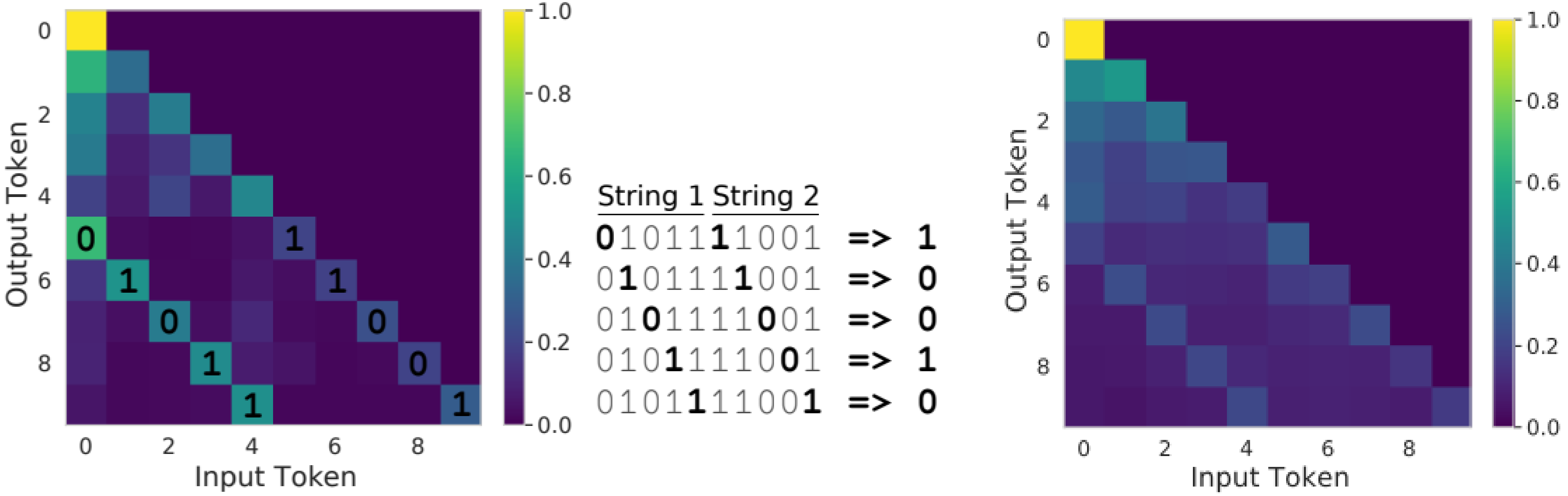


(b) localization

Self-supervised Model as Universal Computation Engine







Initialization	Memory	XOR	ListOps	MNIST	C10	C10 LRA	Homology
Pretrained	100%	100%	38.4%	98.0%	68.2%	38.6%	12.7%
Statistics Only	100%	100%	37.4%	97.2%	56.5%	33.1%	11.0%
Default	75.8%	100%	34.3%	91.7%	61.7%	36.1%	9.3%

這些發現有甚麼用？

Speech Question Answering

TOEFL Listening Comprehension Test by Machine

Link: https://github.com/iamyuanchung/TOEFL-QA

https://arxiv.org/abs/1608.06378

Audio Story: (The original story is 5 min long.)

Question: “ What is a possible origin of Venus’ clouds? ”

Choices:

(A) gases released as a result of volcanic activity

(B) chemical reactions caused by high surface temperatures

(C) bursts of radio energy from the plane's surface

(D) strong winds that blow dust into the atmosphere

https://arxiv.org/abs/1804.00320
https://arxiv.org/abs/1808.02280

SQuAD-style Spoken QA

- Link: https://github.com/chiahsuan156/ODSQA

Dataset	QA-pairs	Hours	M-spkrs	F-spkrs	WER-D(%)	WER-Q(%)	Avg D Len	AvgQ Len
ODSQA	3654	25.28	7	13	19.11	18.57	428	22
DRCD-TTS	16746	--	--	--	33.63	--	332	20

SPOKEN OPEN-DOMAIN QUESTION ANSWERING DATASET

SQuAD-style Spoken QA

- Link: https://github.com/chiahsuan156/ODSQA

Dataset	QA-pairs	Hours	M-spkrs	F-spkrs	WER-D(%)	WER-Q(%)	Avg D Len	AvgQ Len
ODSQA	3654	25.28	7	13	19.11	18.57	428	22
DRCD-TTS	16746	--	--	--	33.63	--	332	20

SPOKEN OPEN-DOMAIN QUESTION ANSWERING DATASET

SOD QA

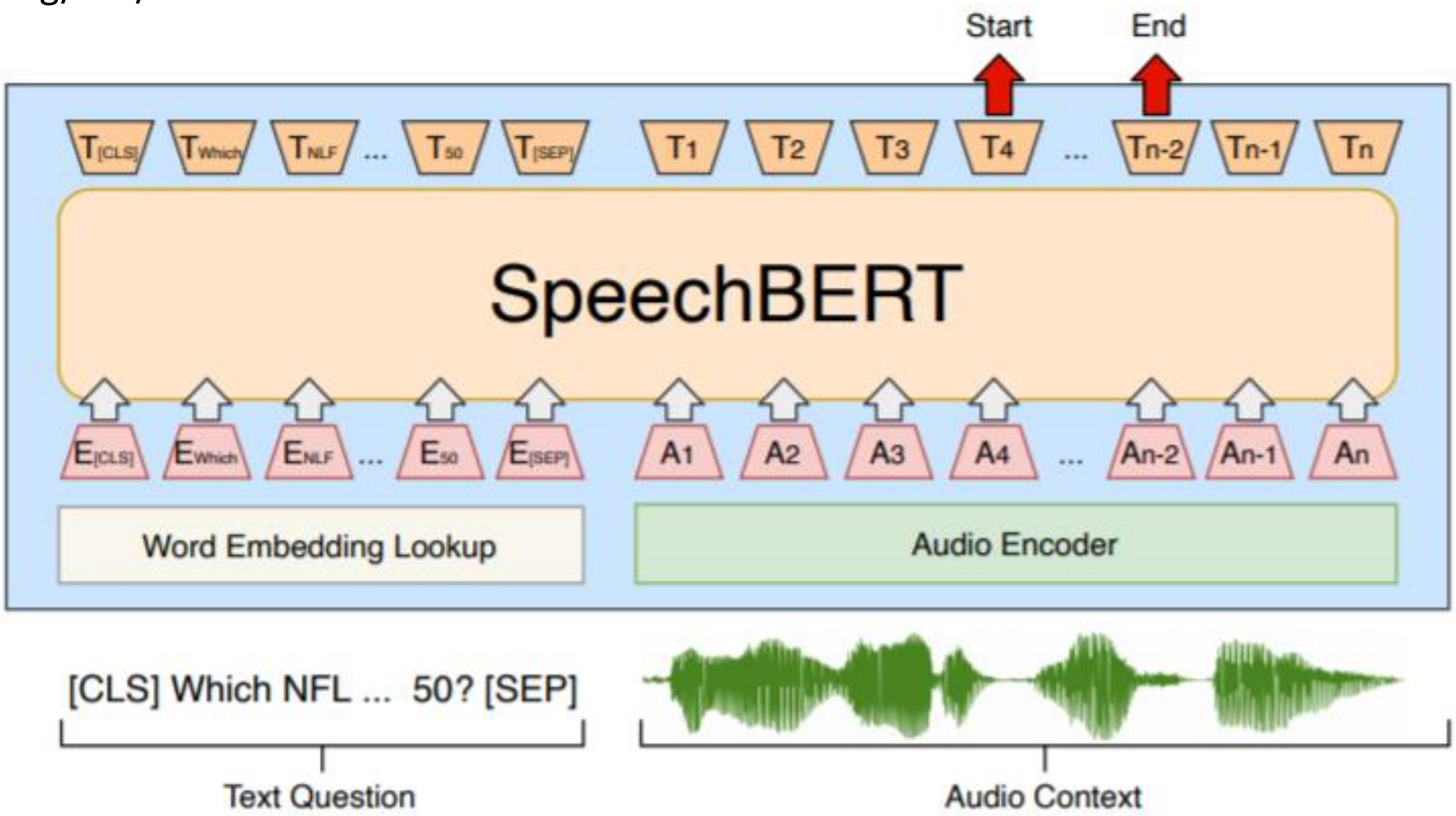
OPEN-DOMAIN SPOKEN QUESTION ANSWERING DATASET

ODS QA

[Lee, et al., SLT'18]

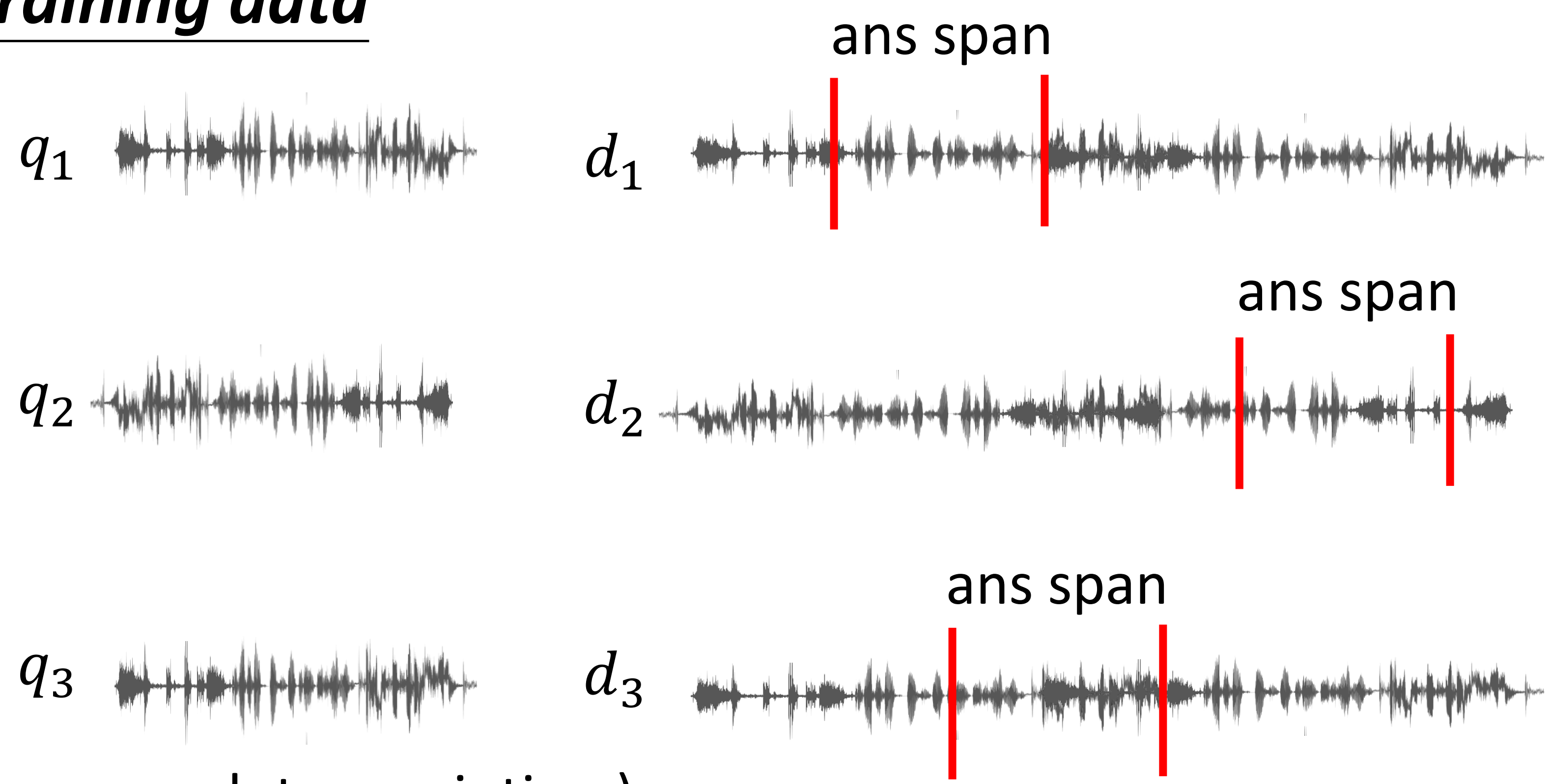
Towards End-to-end

https://arxiv.org/abs/1910.11559



Speech Question Answering

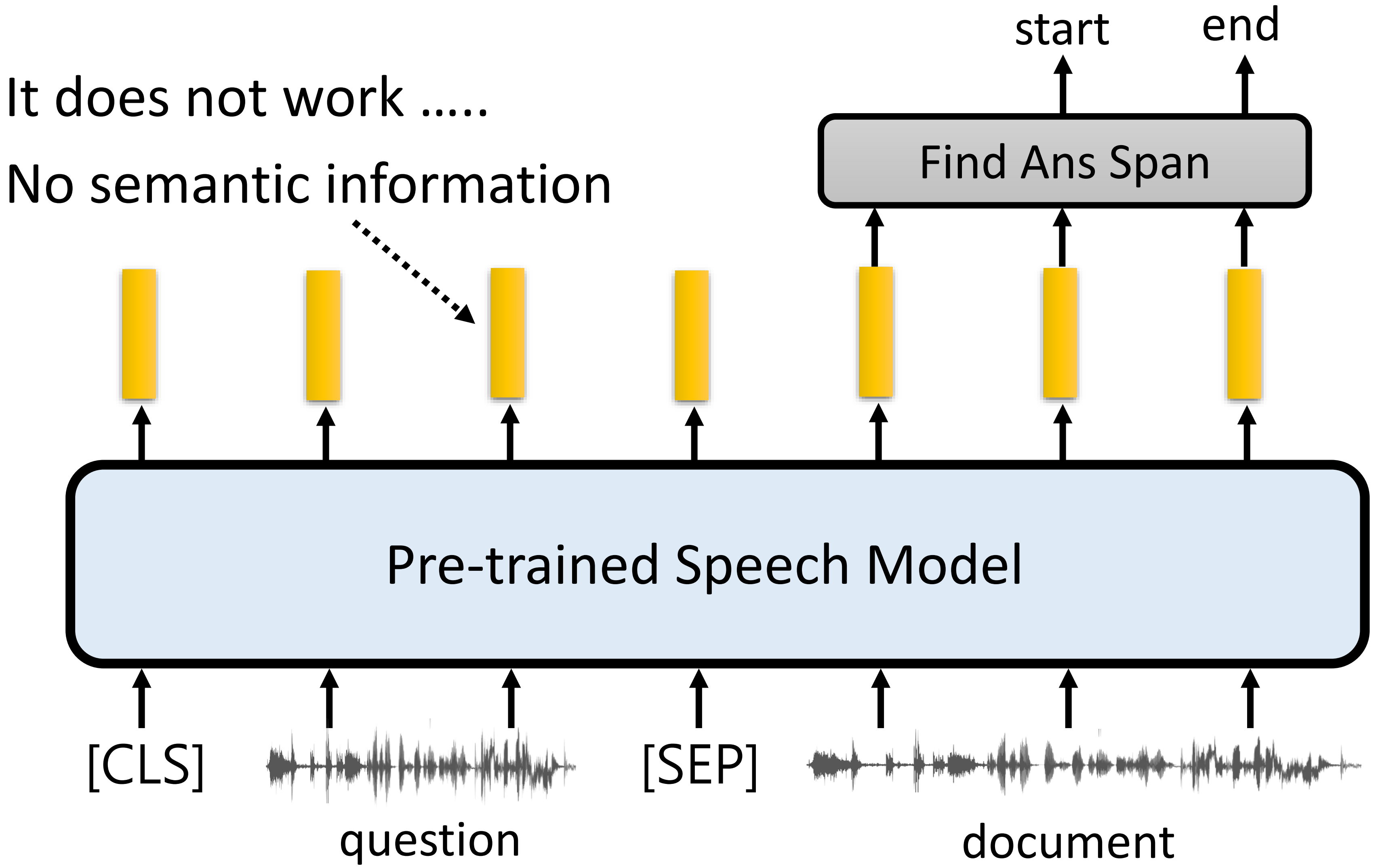
Training data

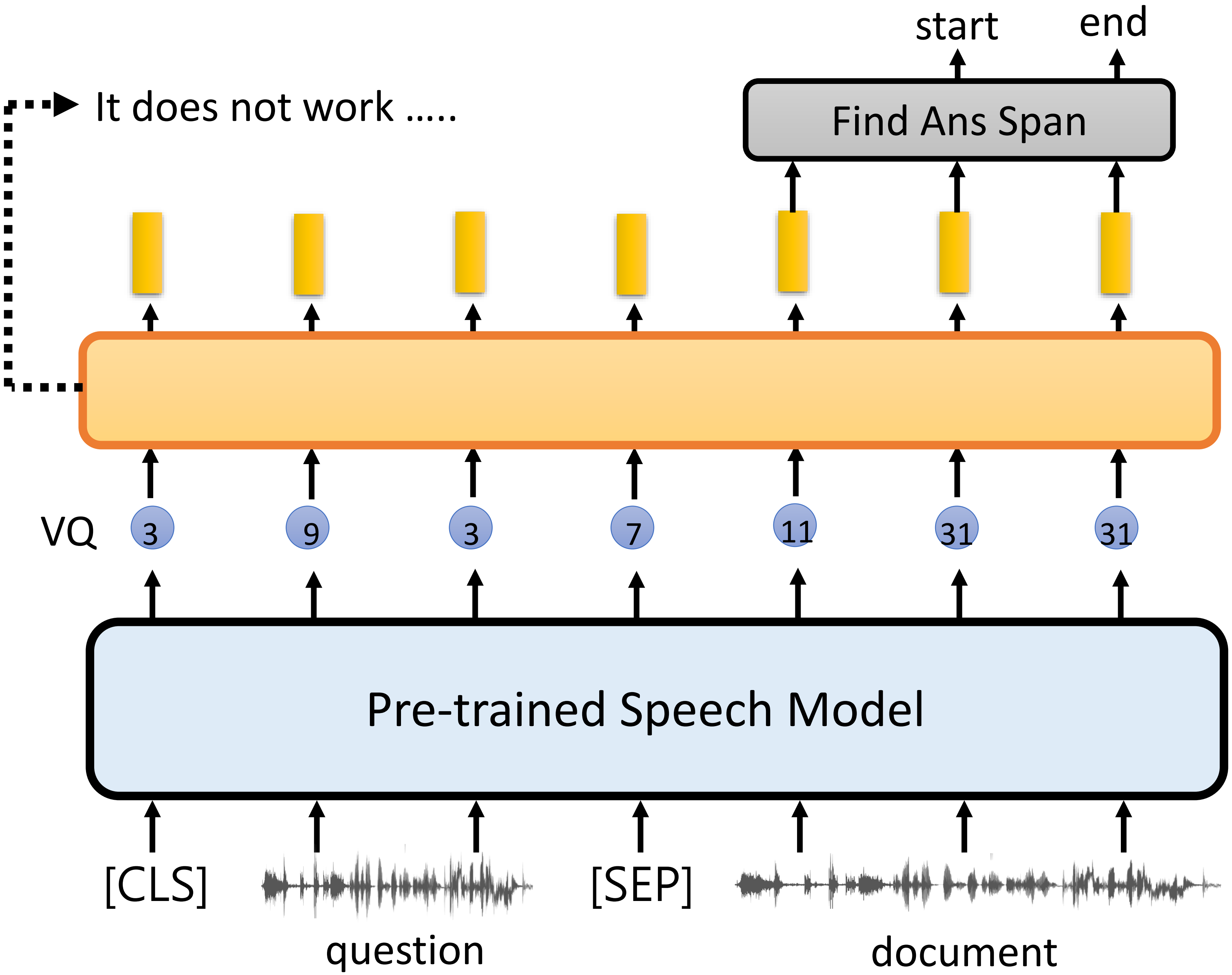


(no speech transcription)

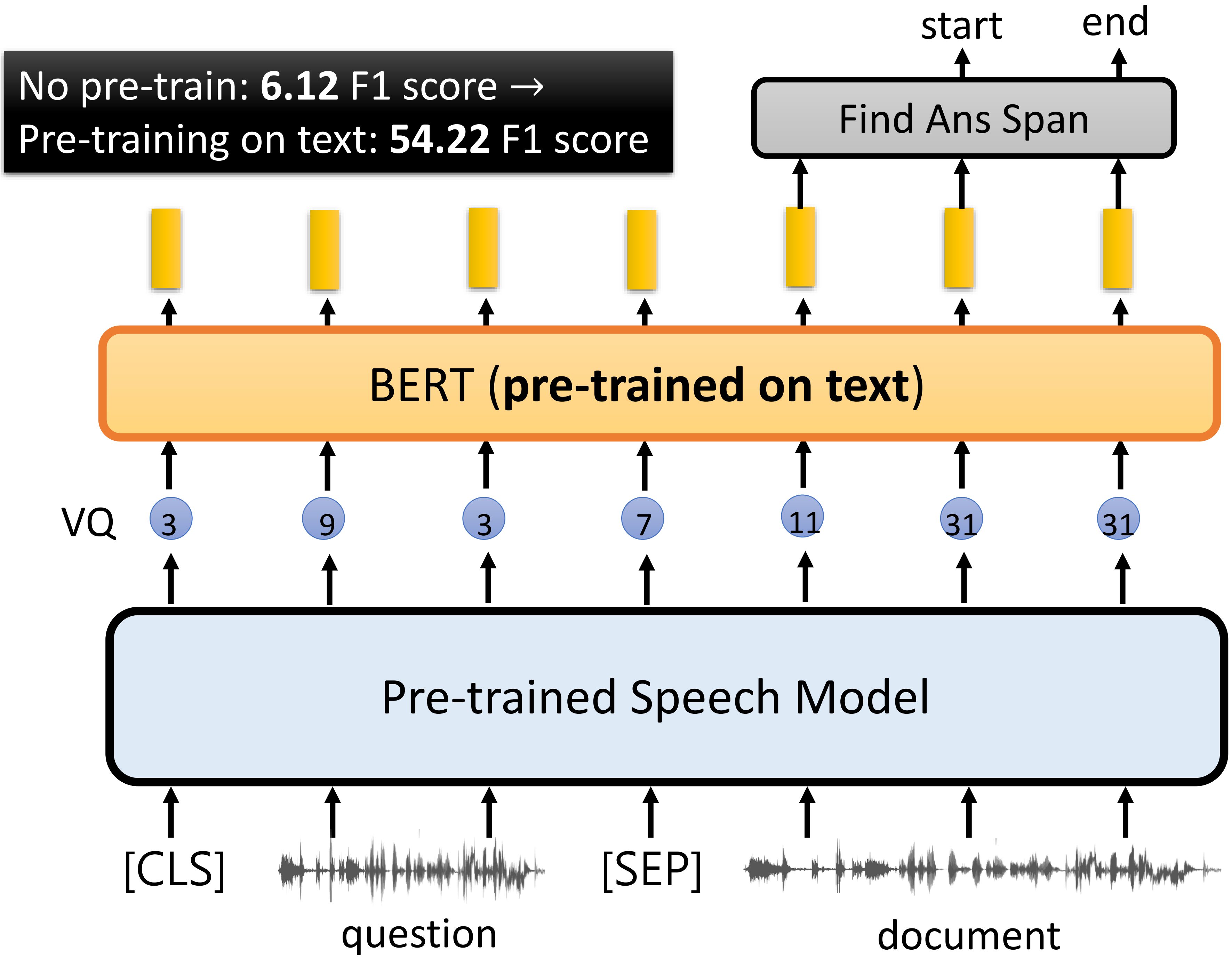
Can we train an end-to-end speech QA model?

Speech Question Answering





start
end
It does not work .....
Find Ans Span
VQ
3
9
3
7
11
31
31
Pre-trained Speech Model
[CLS]
[SEP]
question
document

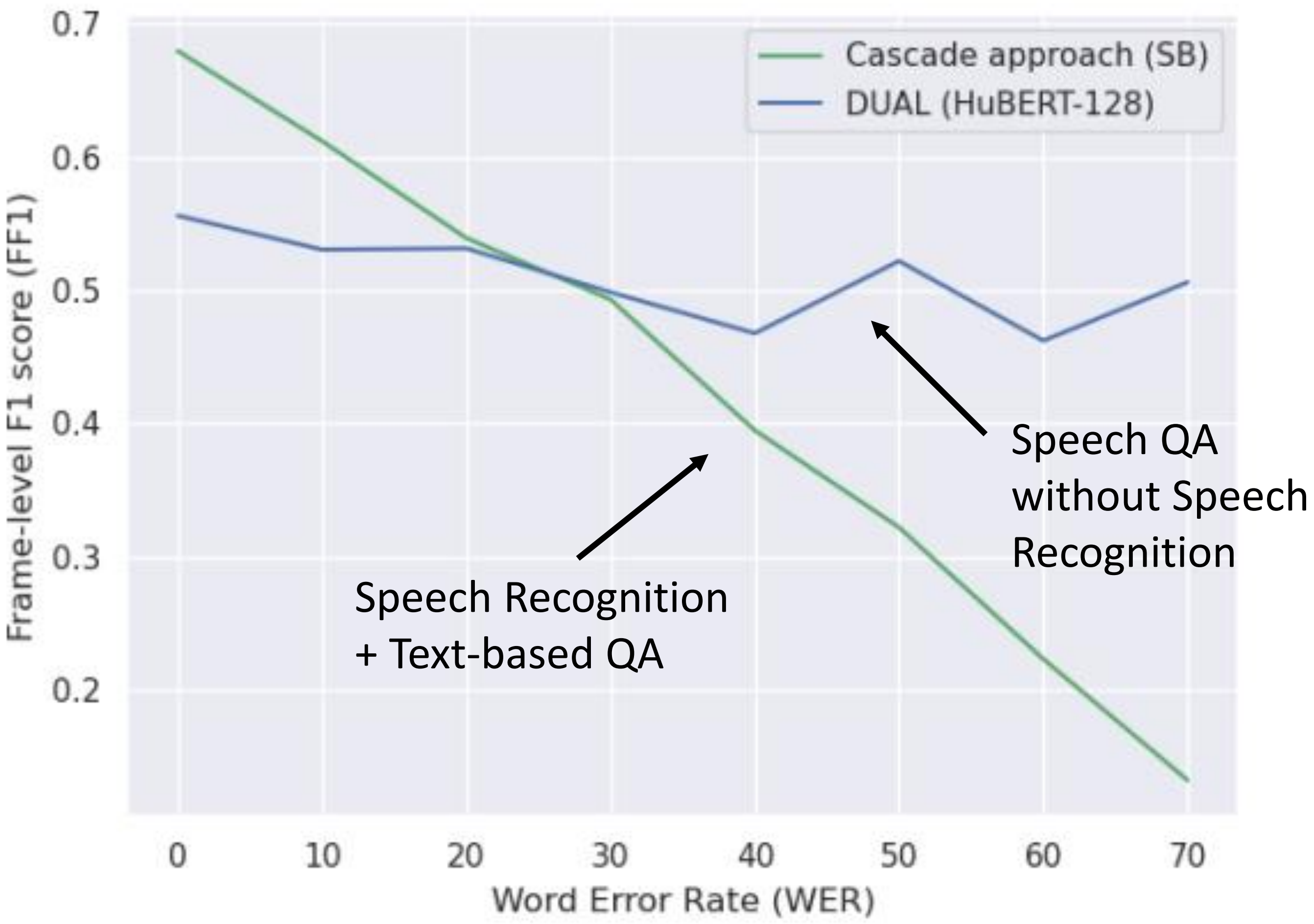

No pre-train: 6.12 F1 score →
Pre-training on text: 54.22 F1 score
start
end
Find Ans Span
BERT (pre-trained on text)
VQ
3
9
3
7
11
31
31
Pre-trained Speech Model
[CLS]
[SEP]
question
document

https://arxiv.org/abs/2203.04911

This work is done by 林冠廷

Speech Question Answering

Embedding Assignment	FF1	AOS
Most frequent	54.2	48.5
Least frequent	46.9	41.7
Random	51.7	46.2
Re-init	8.9	7.2
Scratch (baseline)	6.1	4.9



Outline

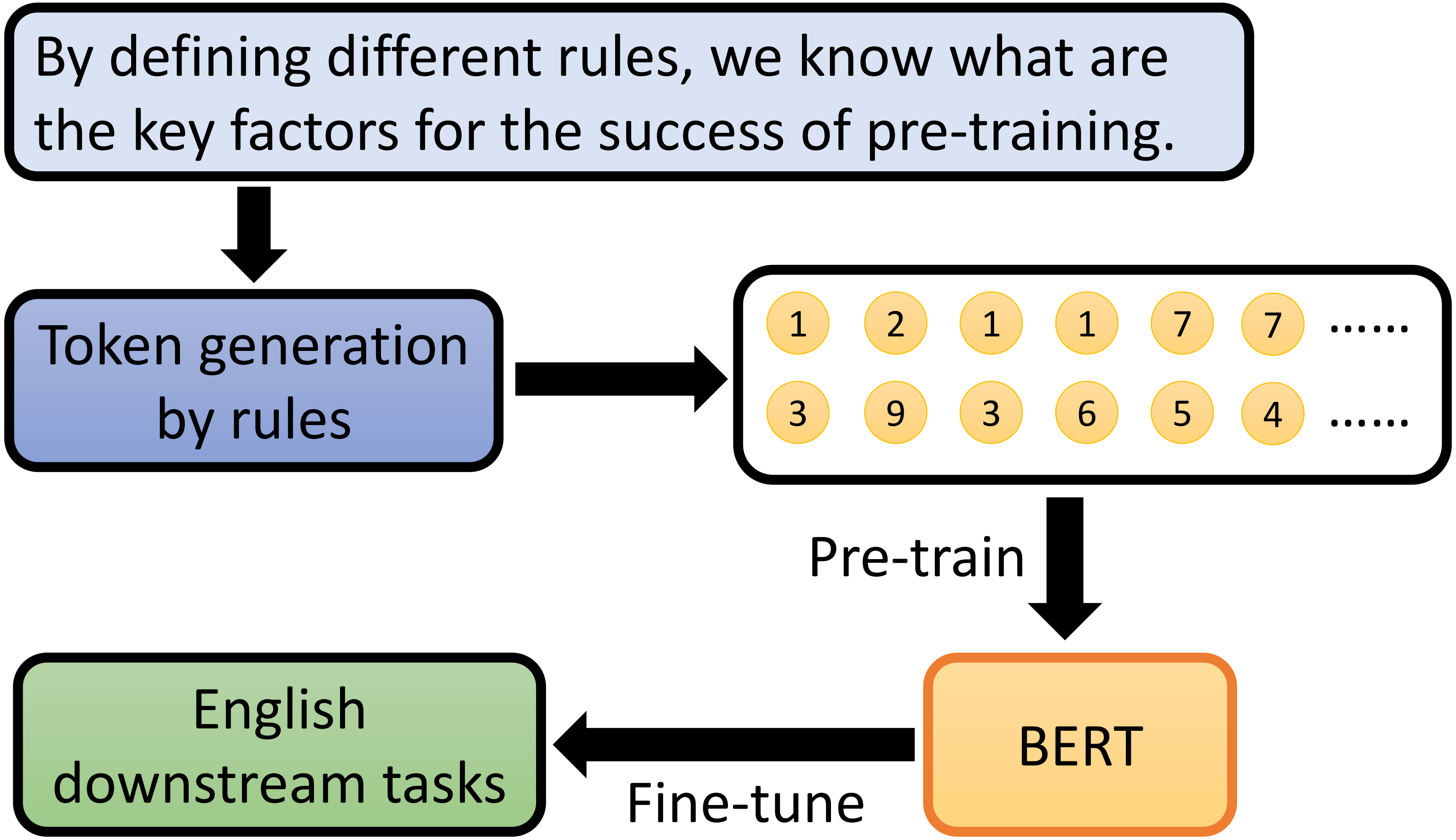
Case Study: BERT

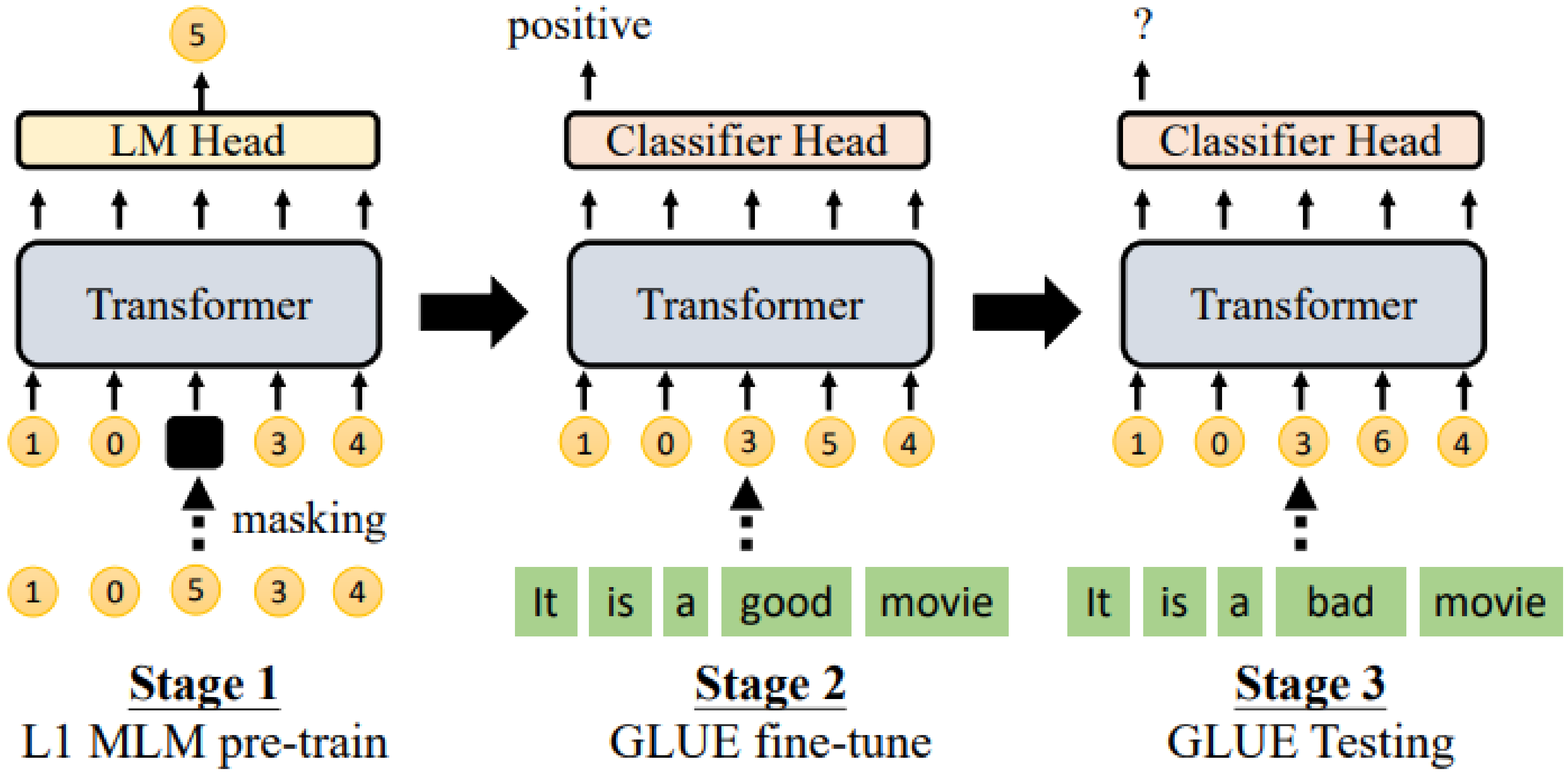
Story 1: Cross-lingual

Story 2: Cross-discipline

Story 3: Pre-training without Human Languages

Pre-training on Artificial Data





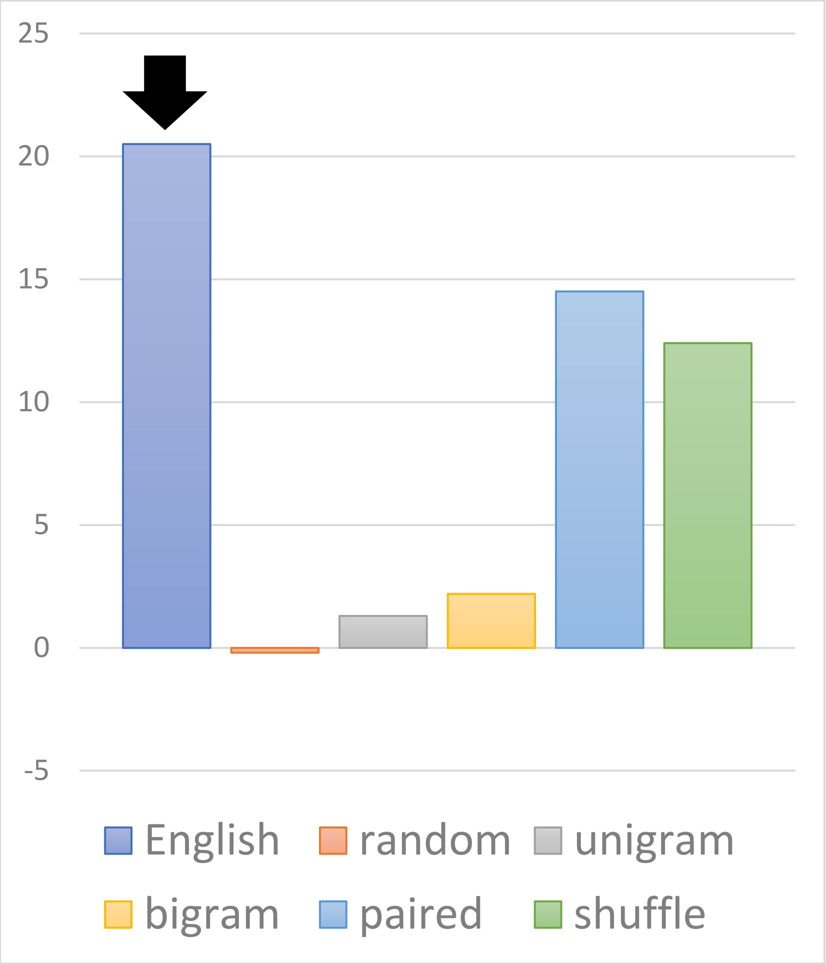
Token ID	Input Token
0	is
1	it
2	very
3	good
4	movie
5	a
6	bad
7	,

https://arxiv.org/abs/2109.03537

This work is done by 姜成翰

Pre-training on Artificial Data

GLUE score improvement vs. train from scratch

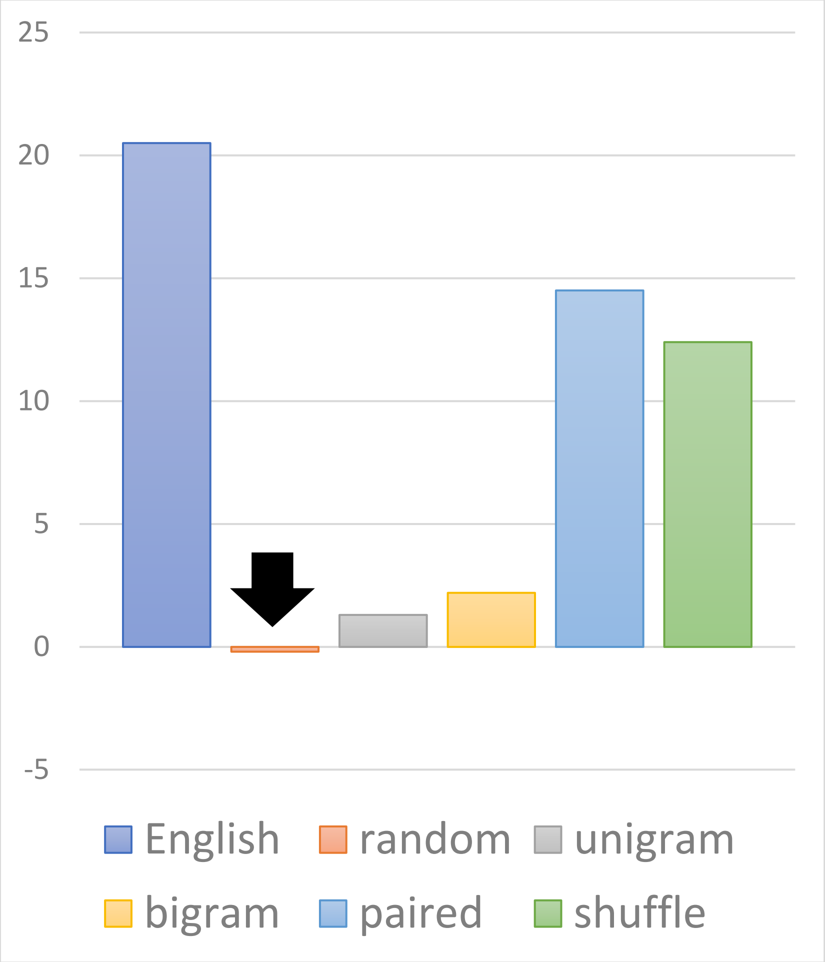


https://arxiv.org/abs/2109.03537

This work is done by 姜成翰

Pre-training on Artificial Data

GLUE score improvement vs. train from scratch



- Pre-training on random generated words yields the same performance as training from scratch.
- Data plays the role.

Learn more from robust wav2vec 2.0

https://arxiv.org/abs/2104.01027

Data is critical ……

Pre-trained Model

HuBERT (Librispeech)

Labeled Data

e.g., clean

Testing Data

e.g., noisy

	continual	IC (Acc) clean	IC (Acc) m+g+r	IC (Acc) fsd50k	ER (Acc) clean	ER (Acc) m+g+r	ER (Acc) fsd50k	KS (Acc) clean	KS (Acc) m+g+r	KS (Acc) fsd50k
(a) baseline	-	99.47	96.94	97.47	63.96	57.33	60.55	97.14	93.38	93.80

	continual	SID (Acc) clean	SID (Acc) m+g+r	SID (Acc) fsd50k	ASR (WER) clean w/o	ASR (WER) clean w/ LM	ASR (WER) m+g+r w/o	ASR (WER) m+g+r w/ LM	ASR (WER) fsd50k w/o	ASR (WER) fsd50k w/ LM	ASR (WER) CHiME3 w/o	ASR (WER) CHiME3 w/ LM
(a) baseline	-	84.97	65.51	77.61	6.72	4.88	10.16	7.94	9.62	7.57	33.4	29.26

Learn more from robust wav2vec 2.0

https://arxiv.org/abs/2104.01027

Data is critical ……

Continuously train with noisy data

Pre-trained Model

HuBERT (Librispeech)

Labeled Data

e.g., clean

Testing Data

e.g., noisy

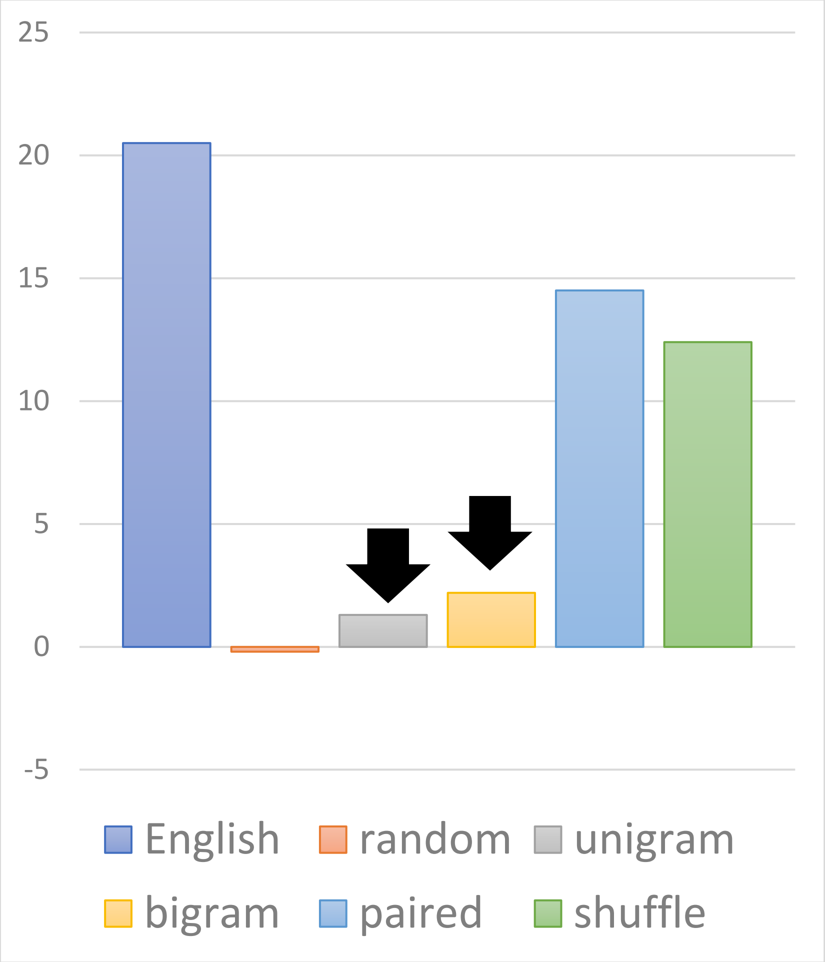
	continual	**IC** (Acc) clean	**IC** (Acc) m+g+r	**IC** (Acc) fsd50k	**ER** (Acc) clean	**ER** (Acc) m+g+r	**ER** (Acc) fsd50k	**KS** (Acc) clean	**KS** (Acc) m+g+r	**KS** (Acc) fsd50k
(a) baseline	-	99.47	96.94	97.47	63.96	57.33	60.55	97.14	93.38	93.80
(c) w/o DAT	libri 100hr mgr	99.45	98.63	97.94	64.42	62.30	60.65	96.92	94.87	93.90
(d) w/o DAT	libri 960hr mgr	99.39	98.84	97.89	67.28	67.47	65.62	97.12	96.11	94.77

	continual	**SID** (Acc) clean	**SID** (Acc) m+g+r	**SID** (Acc) fsd50k	**ASR** (WER) clean w/o	**ASR** (WER) clean w/ LM	**ASR** (WER) m+g+r w/o	**ASR** (WER) m+g+r w/ LM	**ASR** (WER) fsd50k w/o	**ASR** (WER) fsd50k w/ LM	**ASR** (WER) CHiME3 w/o	**ASR** (WER) CHiME3 w/ LM
(a) baseline	-	84.97	65.51	77.61	6.72	4.88	10.16	7.94	9.62	7.57	33.4	29.26
(c) w/o DAT	libri 100hr mgr	87.02	70.91	80.96	6.23	4.87	8.04	6.47	7.90	6.38	27.82	24.27
(d) w/o DAT	libri 960hr mgr	86.40	74.46	81.47	5.92	4.84	7.19	6.00	7.15	5.87	23.83	20.81

https://arxiv.org/abs/2109.03537
This work is done by 姜成翰

Pre-training on Artificial Data

GLUE score improvement vs. train from scratch



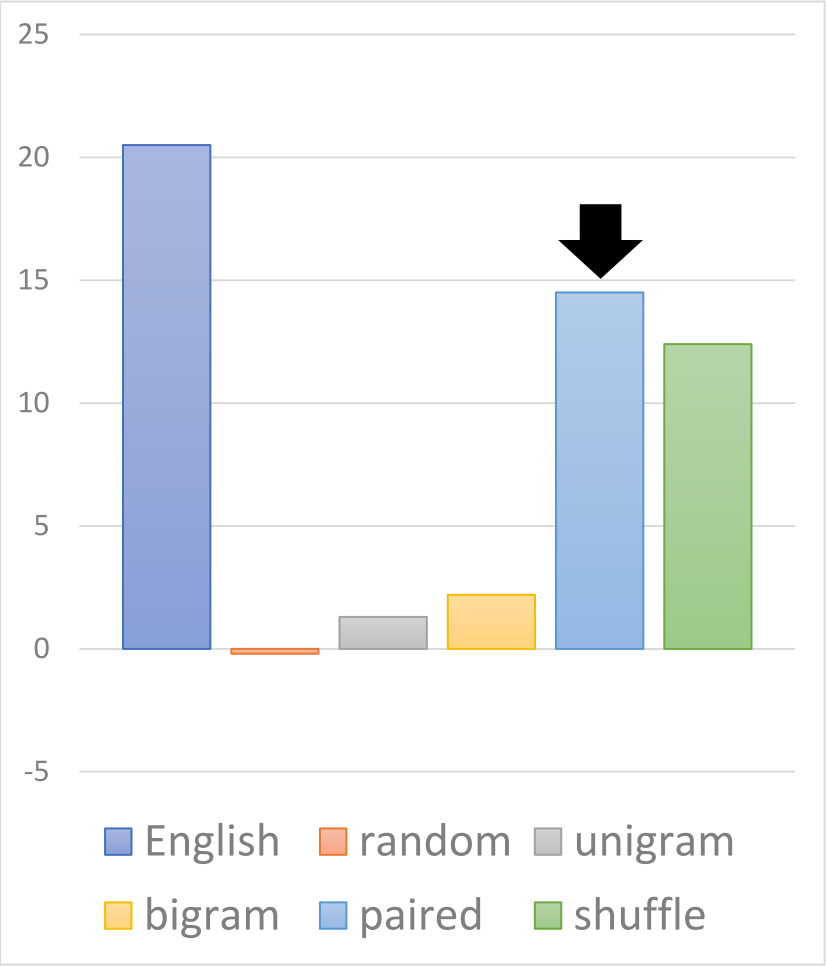
- Pre-training on random generated words yields the same performance as training from scratch.
- Data plays the role.
- Pre-training from the data generated by unigram or bi-gram LMs helps a little.

https://arxiv.org/abs/2109.03537

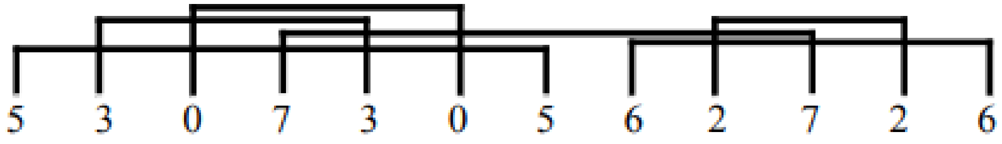
This work is done by 姜成翰

Pre-training on Artificial Data

GLUE score improvement vs. train from scratch



- The sentences generated by very simple rules can lead to good pre-trained models.
- All the words in the generated sentences are paired.



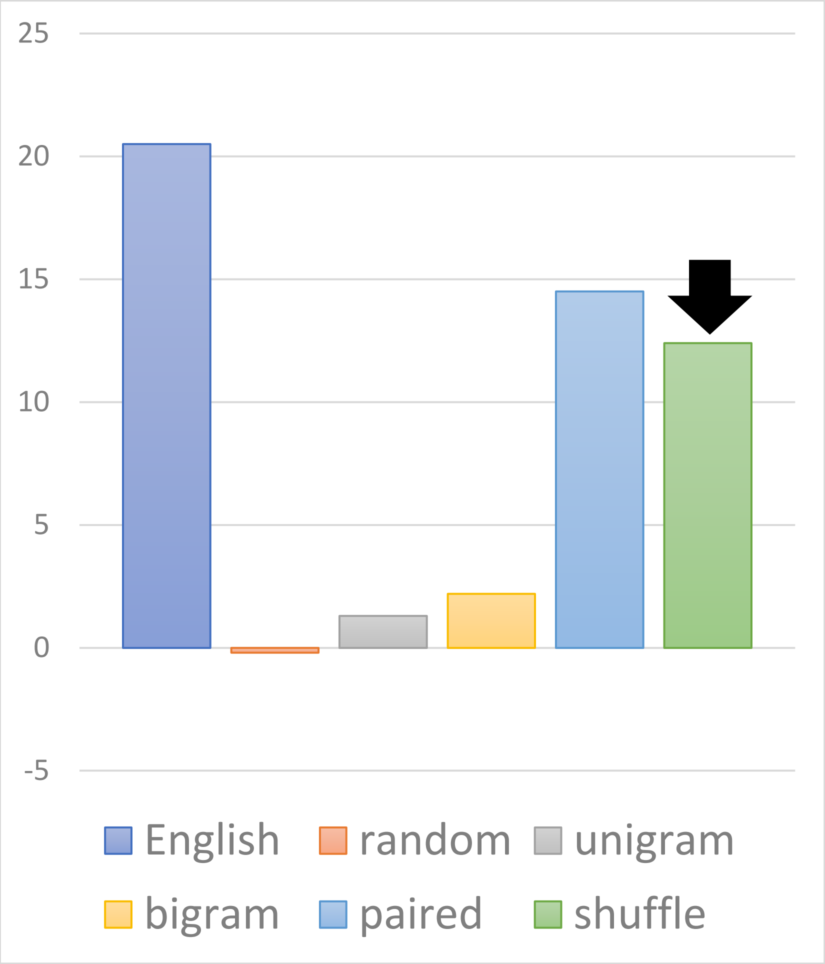
Also refer to:
Learning Music Helps You Read: Using Transfer to Study Linguistic Structure in Language Models

https://arxiv.org/abs/2004.14601

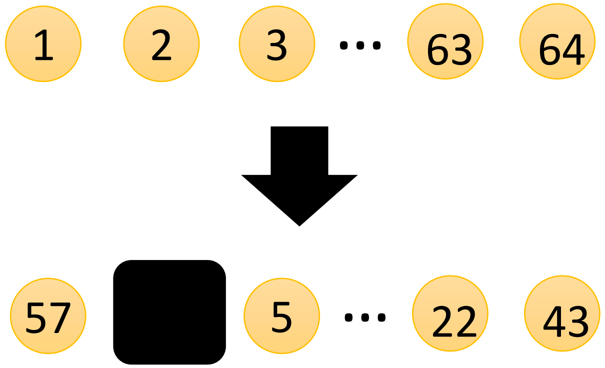
https://arxiv.org/abs/2109.03537
This work is done by 姜成翰

Pre-training on Artificial Data

GLUE score improvement vs. train from scratch



- The sentences generated by very simple rules can lead to good pre-trained models.
- All the words in the generated sentences are paired.
- Shuffle



Concluding Remarks

Story 1: Cross-lingual

Story 2: Cross-discipline

Story 3: Pre-training without Human Languages