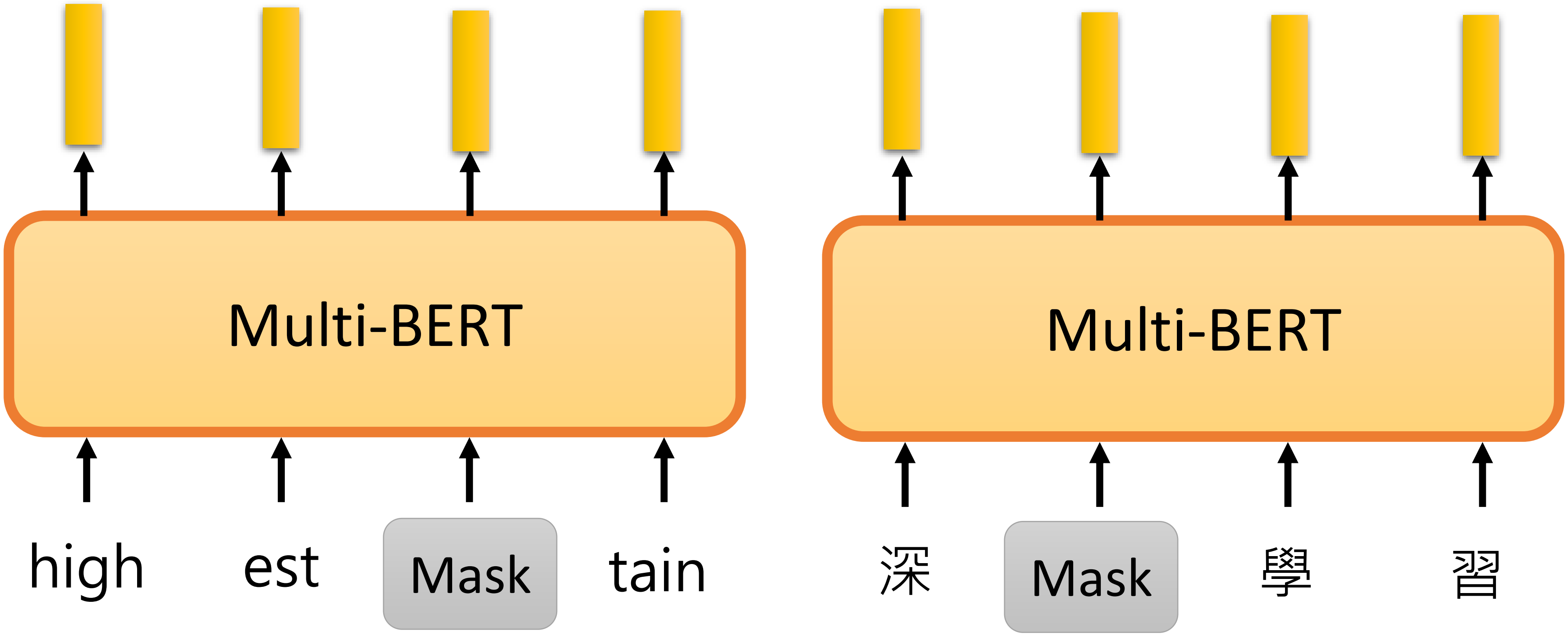
Multilingual BERT

Hung-yi Lee 李宏毅

Source of image: https://www.marstranslation.com/blog/the-sweetest-languages-in-the-world

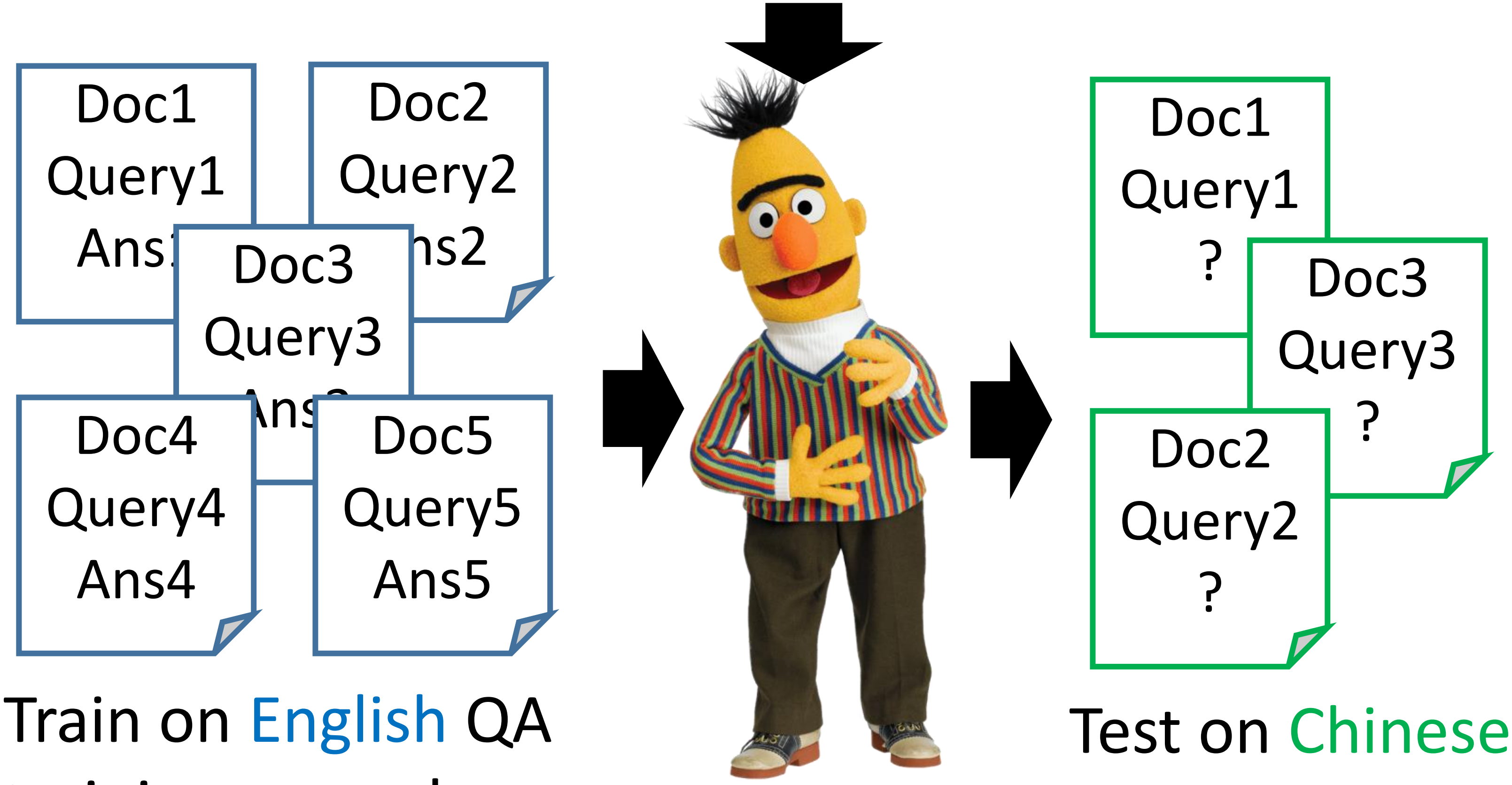
Multi-lingual BERT



Training a BERT model by many different languages.

Zero-shot Reading Comprehension

Training on the sentences of 104 languages



Doc1
Query1
Ans1

Doc2
Query2
Ans2

Doc3
Query3
Ans3

Doc4
Query4
Ans4

Doc5
Query5
Ans5

Doc1
Query1
?

Doc3
Query3
?

Doc2
Query2
?

Train on English QA training examples

Multi-BERT

Test on Chinese QA test

Zero-shot Reading Comprehension

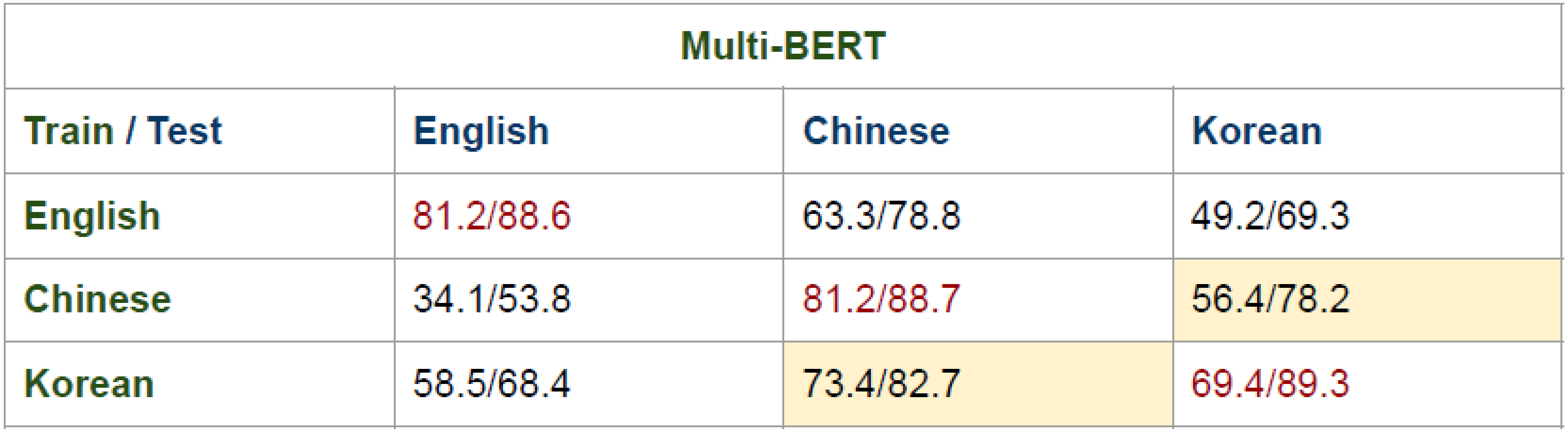
- English: SQuAD, Chinese: DRCD

[Hsu, Liu, et al., EMNLP'19]

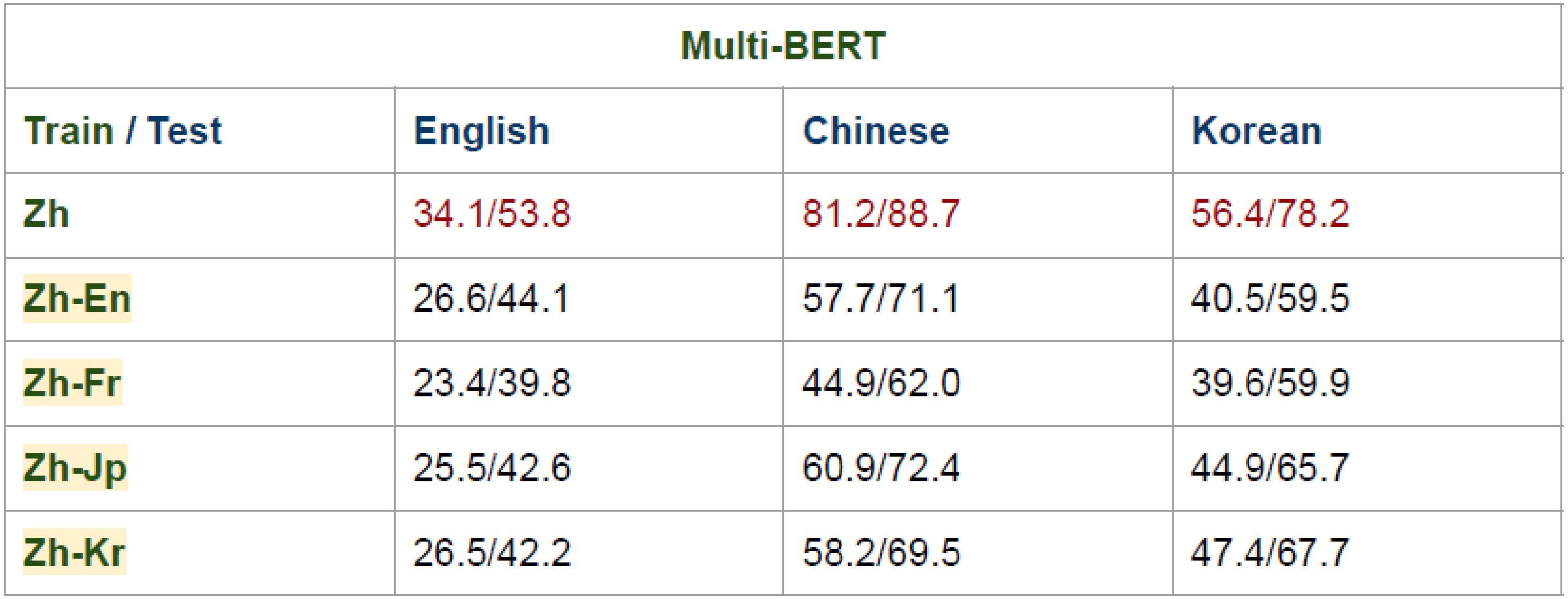
Model	Pre-train	Fine-tune	Test	EM	F1
QANet	none	Chinese	Chinese	66.1	78.1
BERT	Chinese	Chinese		82.0	89.1
	104 languages	Chinese		81.2	88.7
		English		63.3	78.8
		Chinese + English		82.6	90.1

F1 score of Human performance is 93.30%

This work is done by 劉記良、許宗嫄



Multi-BERT			
Train / Test	English	Chinese	Korean
English	81.2/88.6	63.3/78.8	49.2/69.3
Chinese	34.1/53.8	81.2/88.7	56.4/78.2
Korean	58.5/68.4	73.4/82.7	69.4/89.3



Multi-BERT			
Train / Test	English	Chinese	Korean
Zh	34.1/53.8	81.2/88.7	56.4/78.2
Zh-En	26.6/44.1	57.7/71.1	40.5/59.5
Zh-Fr	23.4/39.8	44.9/62.0	39.6/59.9
Zh-Jp	25.5/42.6	60.9/72.4	44.9/65.7
Zh-Kr	26.5/42.2	58.2/69.5	47.4/67.7

[Hsu, Liu, et al., EMNLP’19]

This work is done by 劉記良、許宗嫄

So many evidences ……

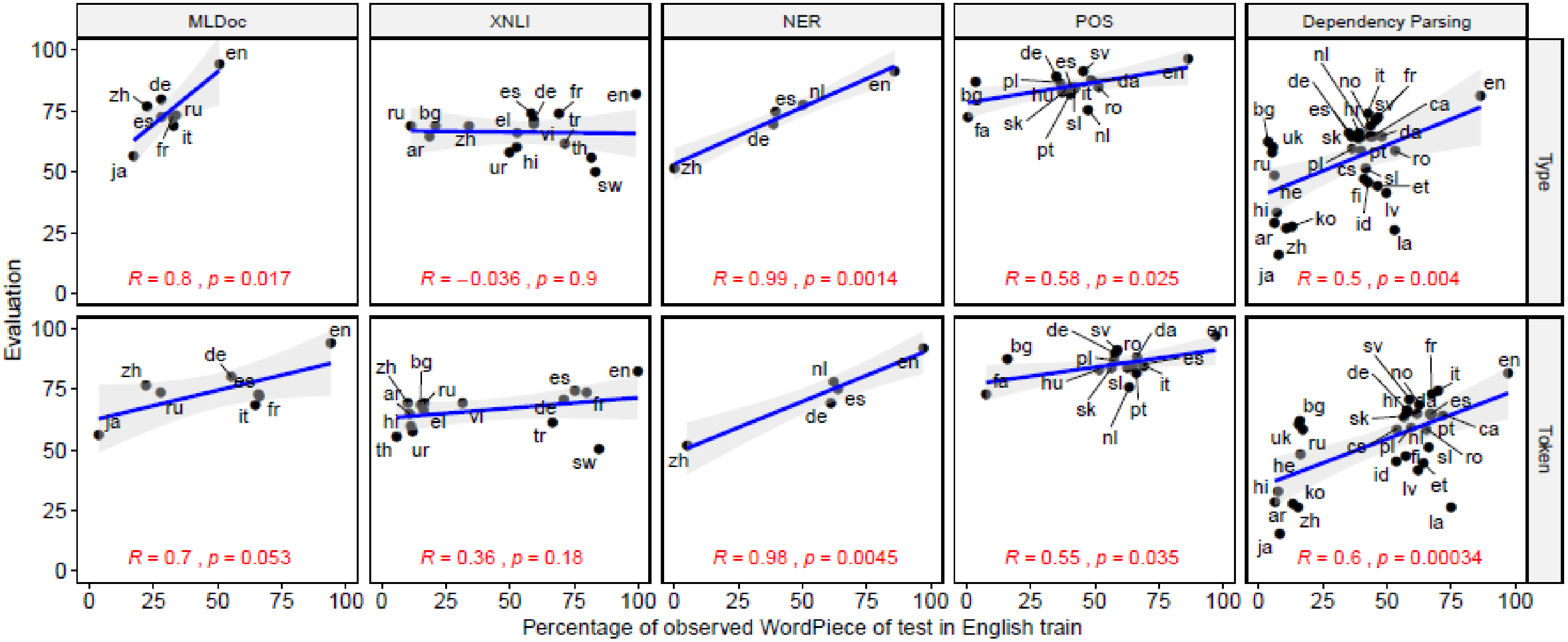
Fine-tuning \ Eval	EN	DE	NL	ES
EN	**90.70**	69.74	77.36	73.59
DE	73.83	**82.00**	76.25	70.03
NL	65.46	65.68	**89.86**	72.10
ES	65.38	59.40	64.39	**87.18**

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.

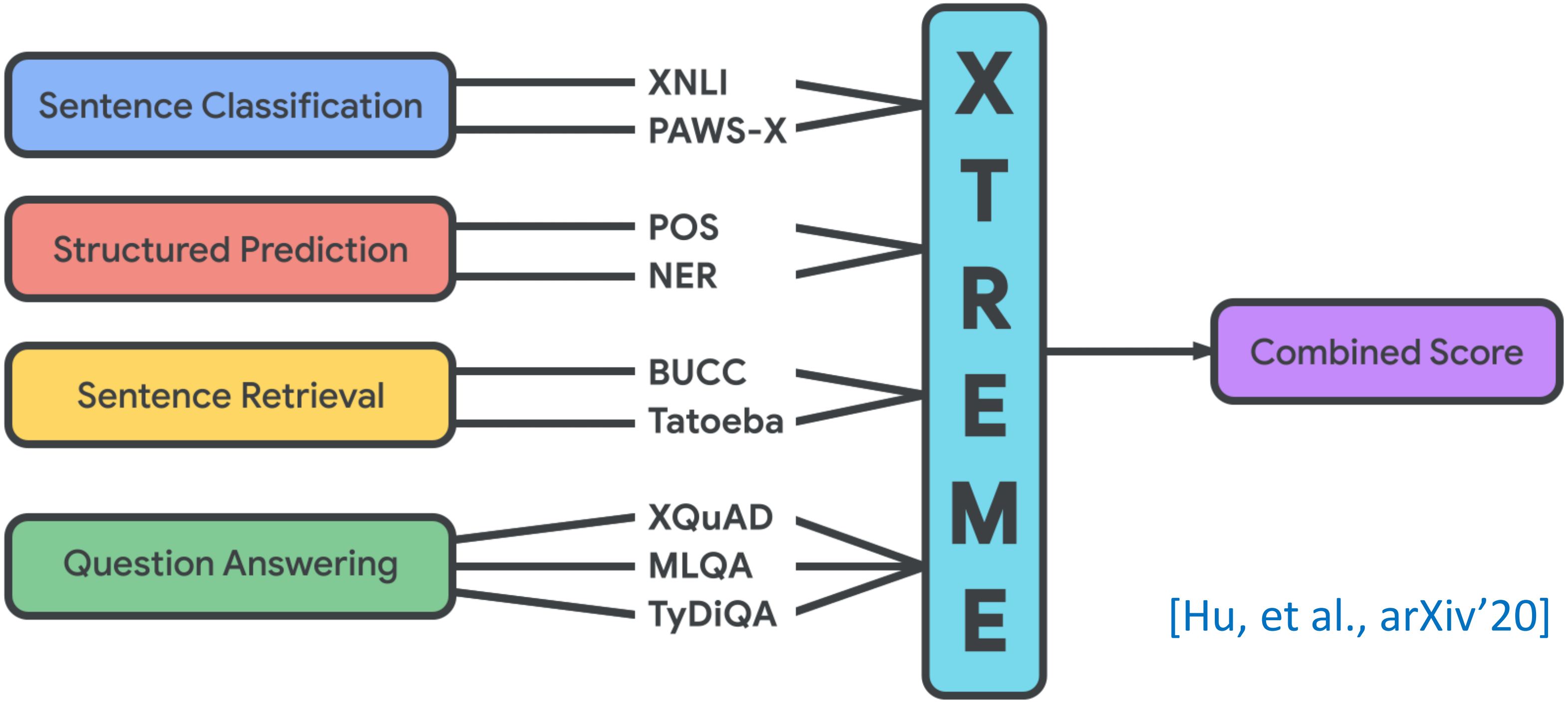
[Pires, et al., ACL'19]



[Wu, et al., EMNLP'19]

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

https://sites.research.google/xtreme

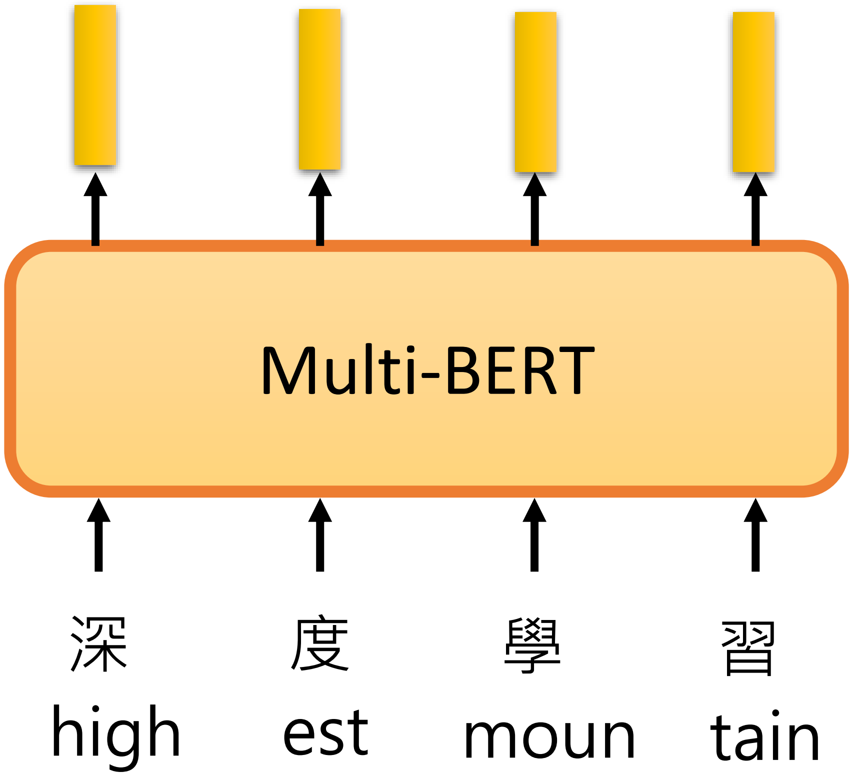


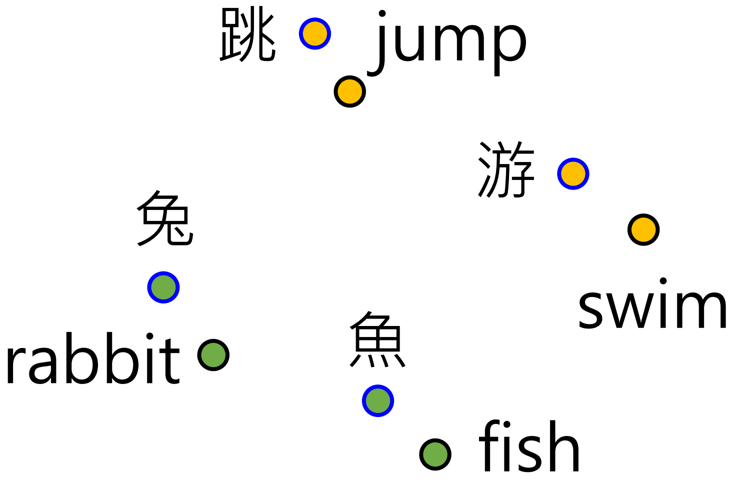
[Hu, et al., arXiv'20]

40 languages for 9 tasks

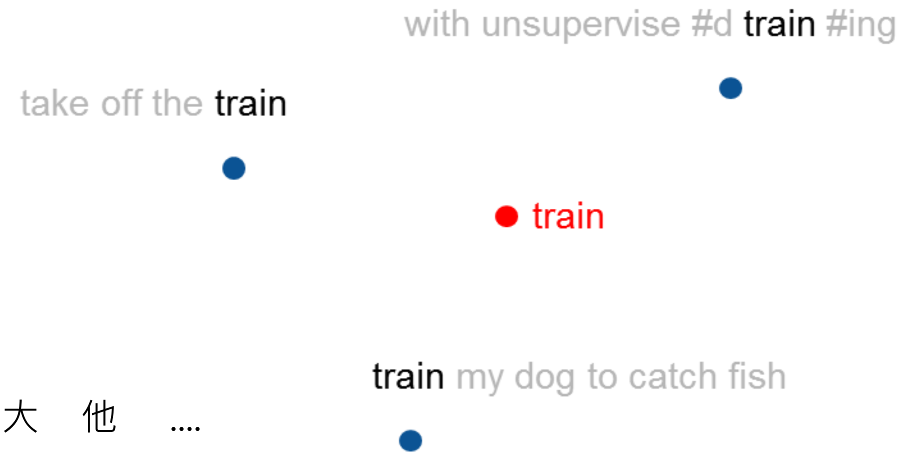
Train on English, and test on the rest

Cross-lingual Alignment?





Mean Reciprocal Rank



	年	月	和	村	人	大	他	....
year	0.7	0.6	0.2	0.1	0.5	0.3	0.4	
month								
and								
village								
man								
big								
he								
⋮								

Cosine Similarity of Representation Vector

投影片來源: 許宗嫄同學碩士口試投影片

Mean Reciprocal Rank

Bi-lingual Dictionary

year →	年
month →	月
village →	村
big →	大

	年	月	和	村	人	大	他	....	Rank	Score
year	0.7	0.6	0.2	0.1	0.5	0.3	0.4		1	1/1
month										
and										
village										
man										
big										
he										
⋮										

投影片來源: 許宗嫄同學碩士口試投影片

Mean Reciprocal Rank

Bi-lingual Dictionary

year → 年
month → 月
village → 村
big → 大

	年	月	和	村	人	大	他	....	Rank	Score
year	0.7	0.6	0.2	0.1	0.5	0.3	0.4		1	1/1
month	0.5	0.6	0.7	0.8	0.1	0.2	0.3		3	1/3
and										
village										
man										
big										
he										
⋮										

投影片來源: 許宗嫄同學碩士口試投影片

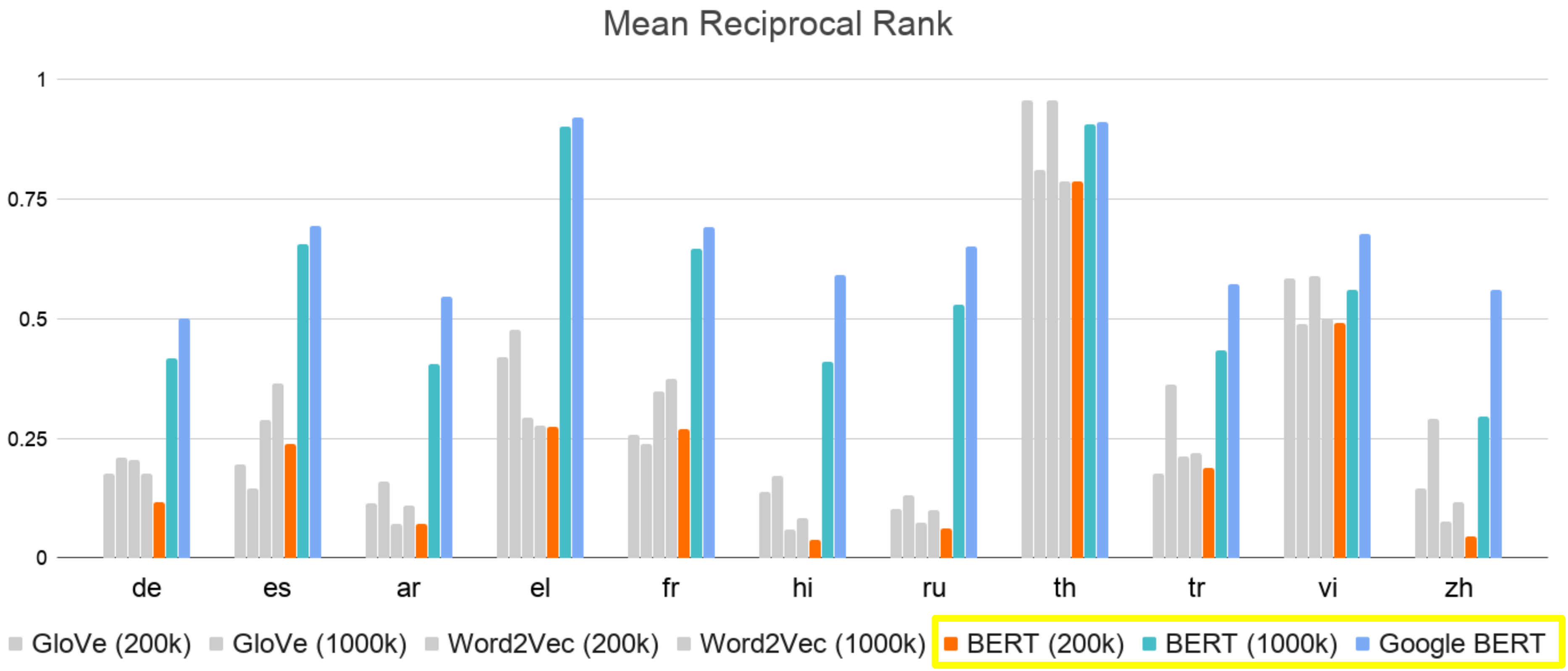
Mean Reciprocal Rank

	年	月	和	村	人	大	他	....	Rank	Score
year	0.7	0.6	0.2	0.1	0.5	0.3	0.4		1	1/1
month	0.5	0.6	0.7	0.8	0.1	0.2	0.3		3	1/3
and	0.1	0.3	0.6	0.5	0.7	0.2	0.4		2	1/2
village	0.5	0.8	0.7	0.6	0.1	0.3	0.1		3	1/3
man	0.1	0.7	0.8	0.6	0.4	0.2	0.3		4	1/4
big	0.3	0.1	0.5	0.8	0.7	0.9	0.2		1	1/1
he	0.5	0.8	0.3	0.6	0.9	0.4	0.7		3	1/3
⋮										

Average

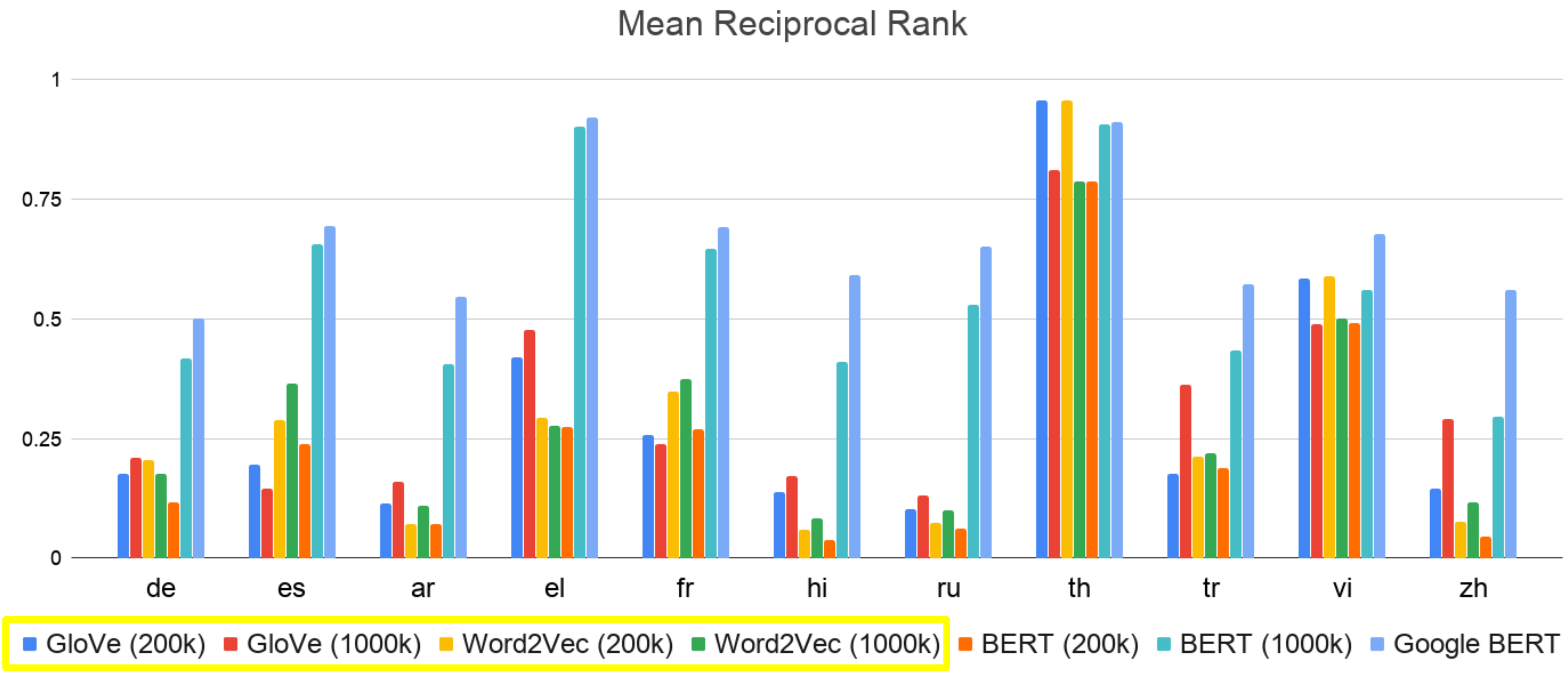
投影片來源: 許宗嫄同學碩士口試投影片

The amount of training data is critical for alignment.



投影片來源: 許宗嫄同學碩士口試投影片

Word2vec and GloVe cannot align well even with more data.



投影片來源: 許宗嫄同學碩士口試投影片

接下來課程規劃

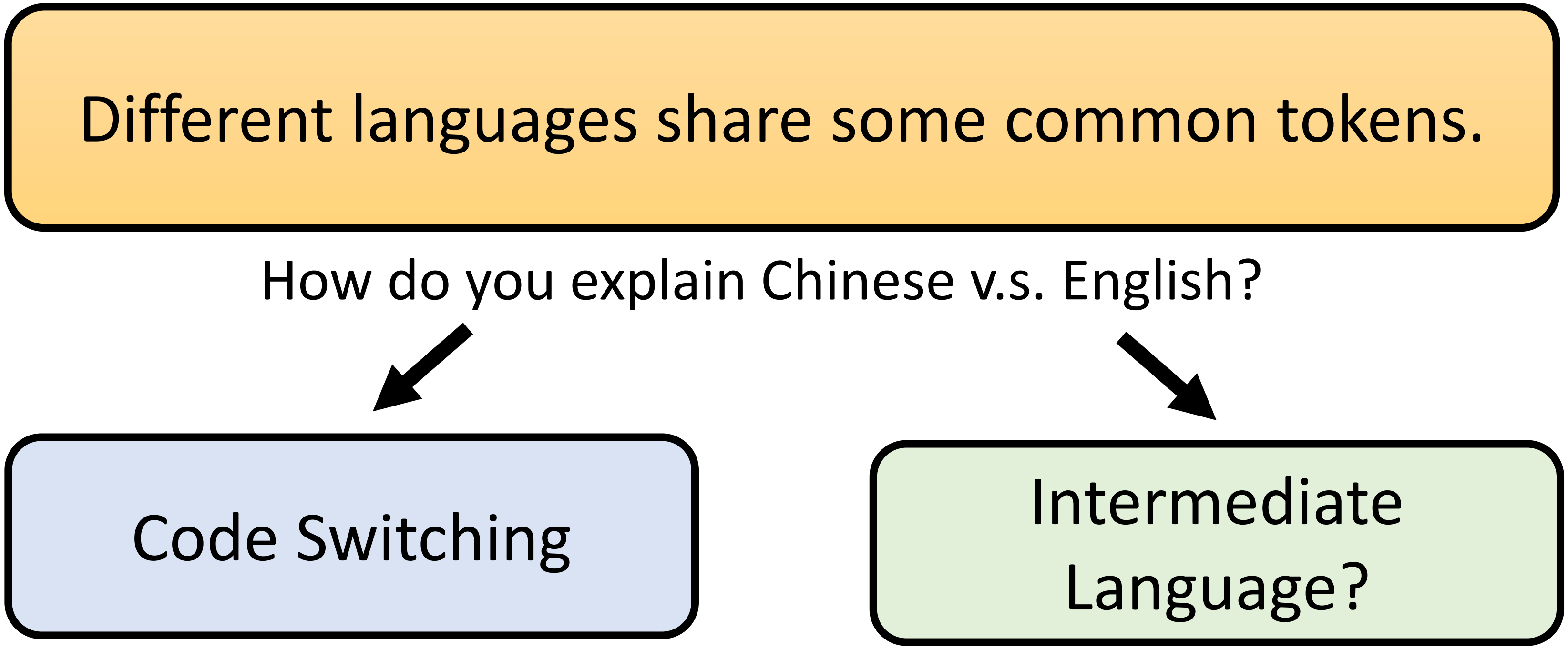
- 接下來的課程都跟作業沒有關係

- 6/17: Multilingual BERT, Dependency Parsing, QA (Part 1)
- 6/24: QA (Part 2), Dialogue State Tracking (as QA), Conditional Sentence Generation
- 7/01: Knowledge graph extraction

- Meta learning / Attacking / Explainable AI for Human Language Processing 暑假必定找時間錄影，決不食言

How alignment happens?

- Typical answer



Different languages share some common tokens.

How do you explain Chinese v.s. English?

Code Switching

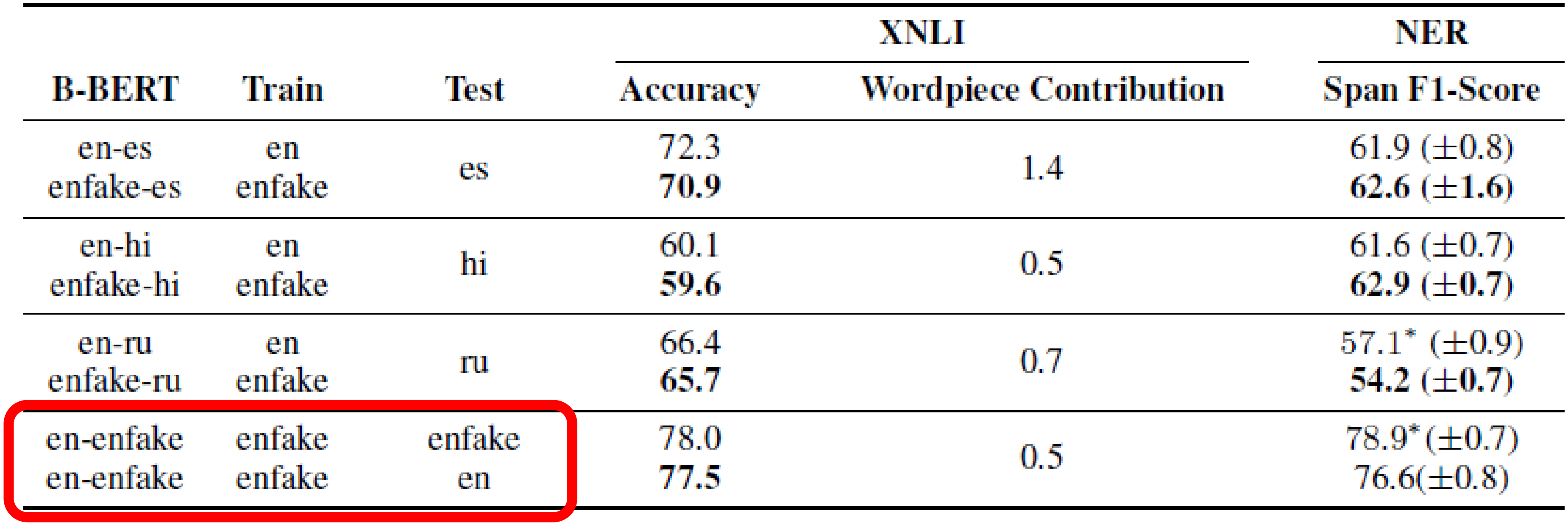
Intermediate Language?

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

Language X shares tokens with Chinese and English.

How alignment happens?

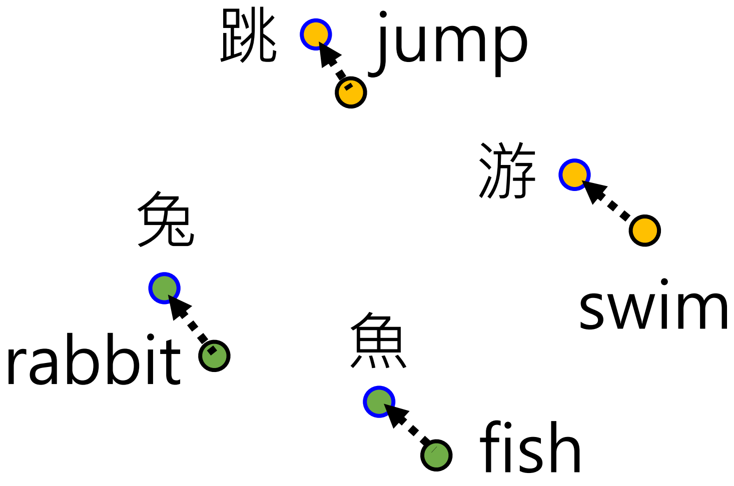
[K, et al., ICLR’20]



			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

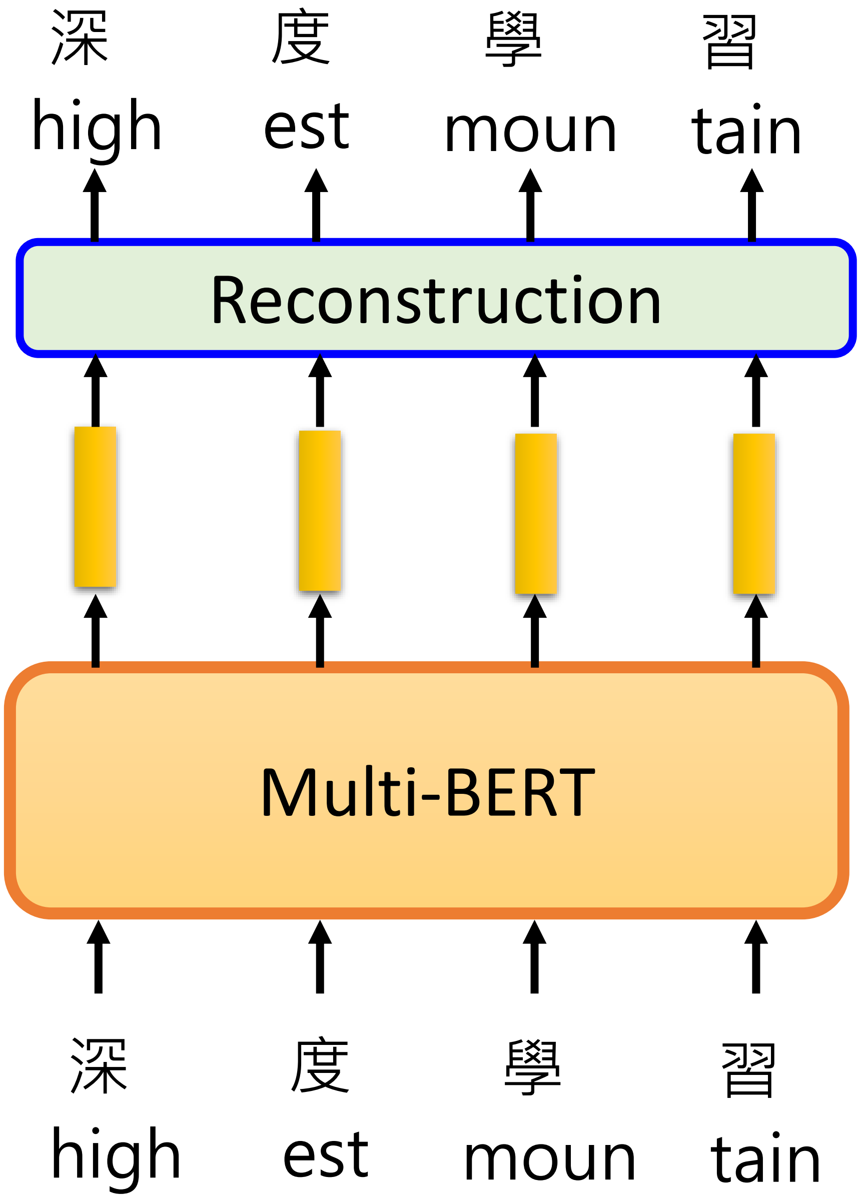
Sounds weird?

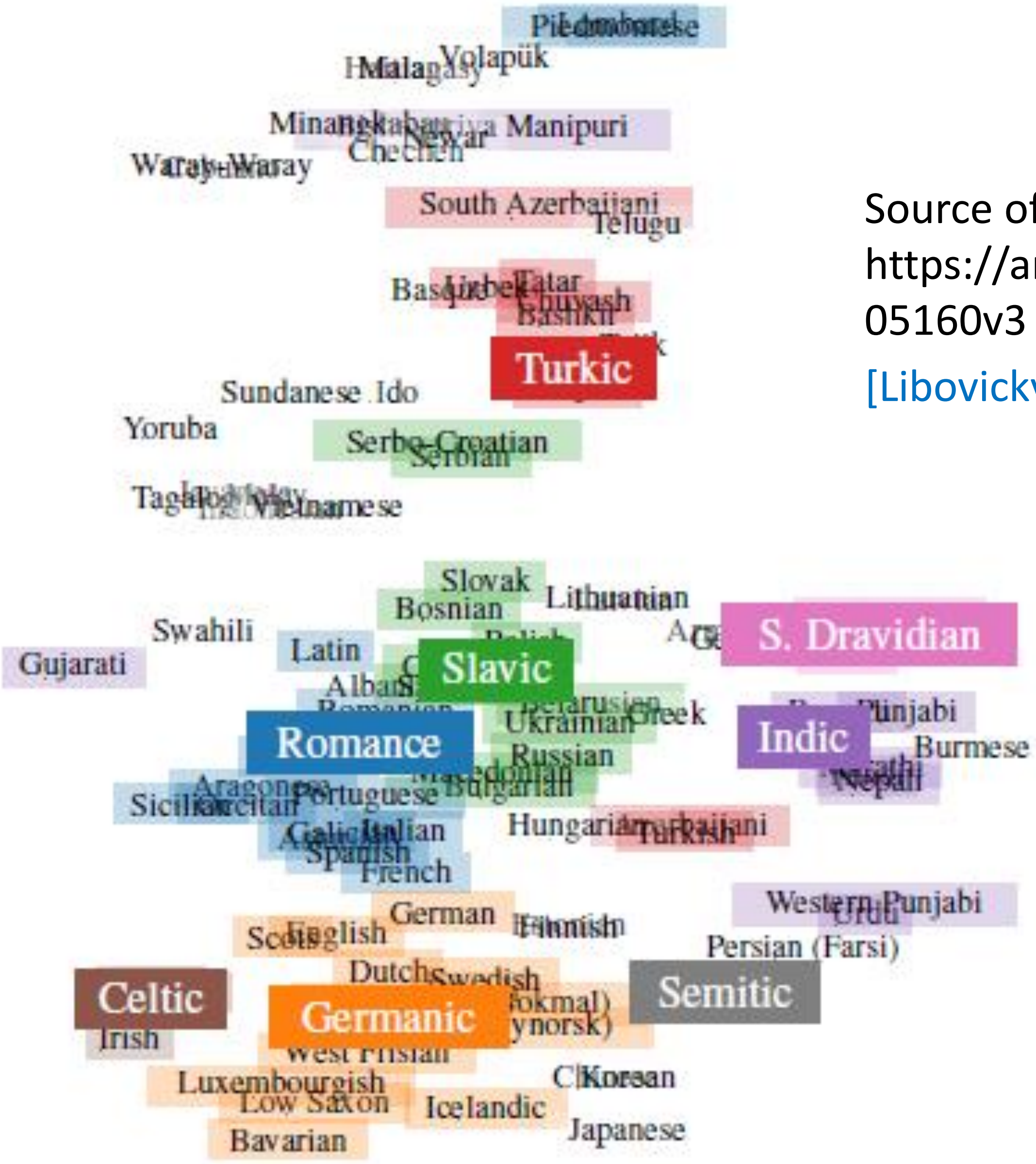


If the embedding is language independent …

How to correctly reconstruct?

There must be language information.



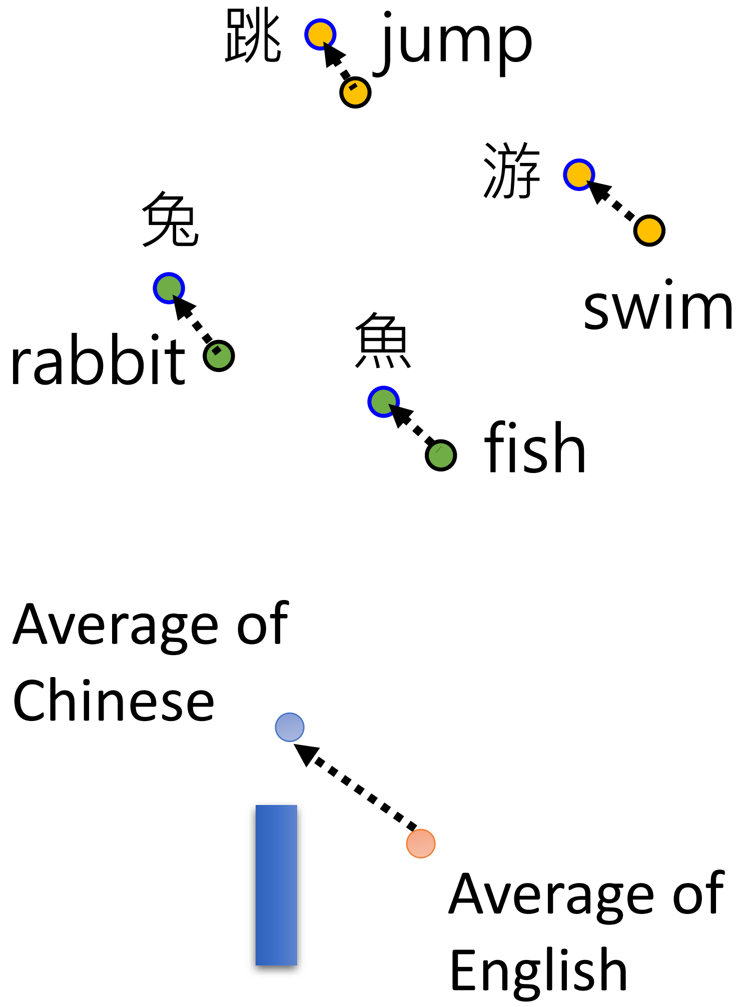


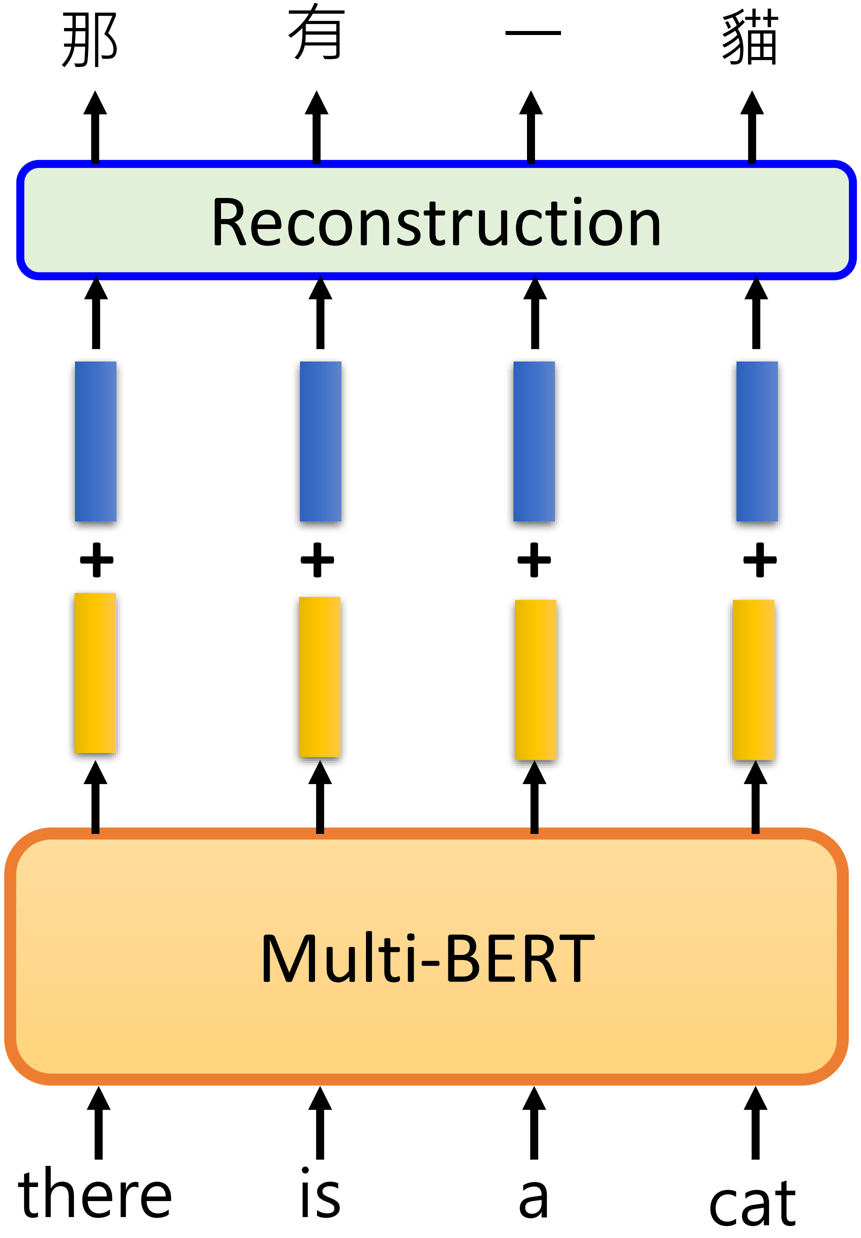
Source of image: https://arxiv.org/abs/2004.05160v3

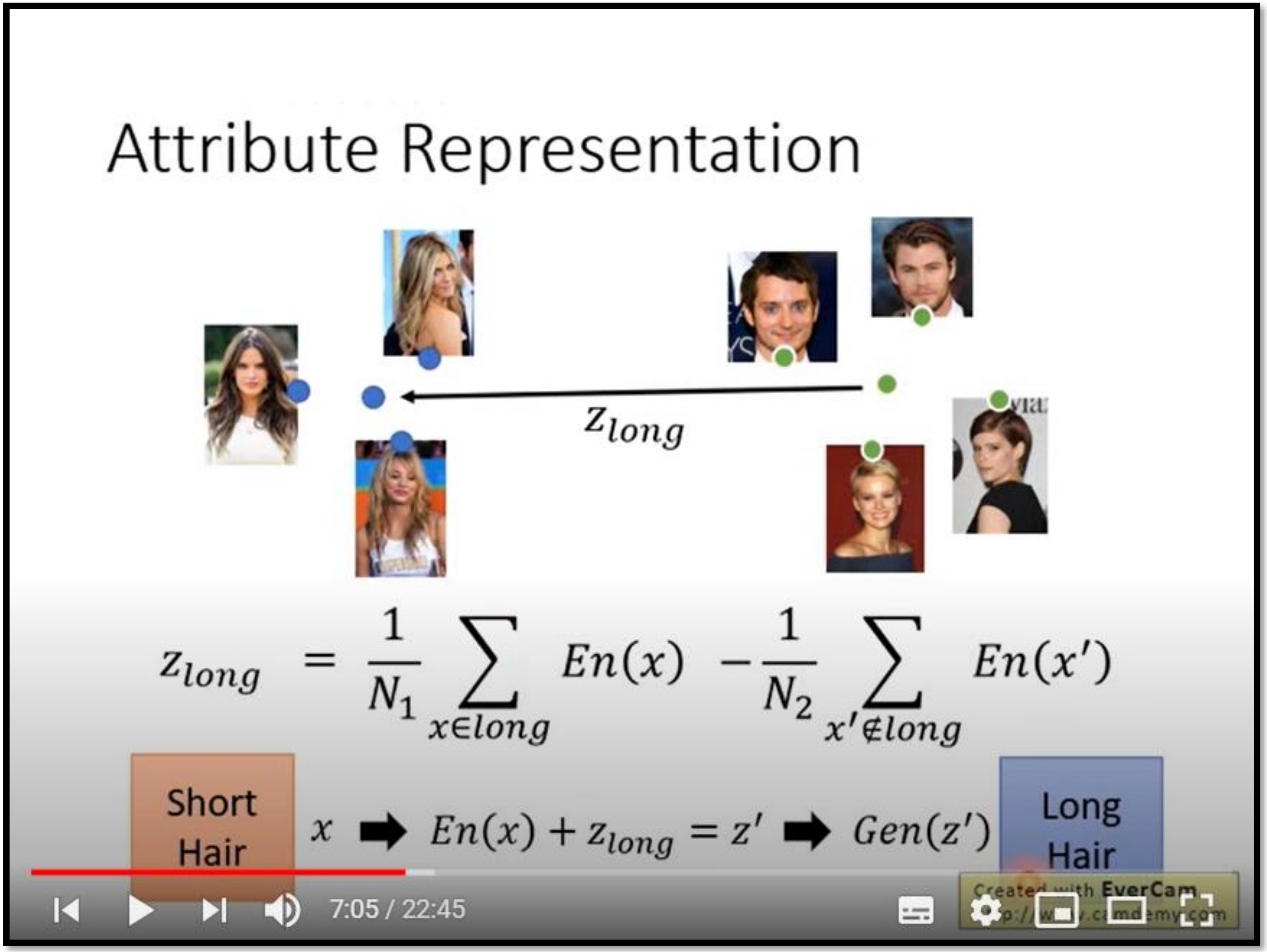
[Libovický, arXiv’20]

This work is done by 劉記良、許宗嫄、莊永松

If this is true …







Attribute Representation

z_{long}

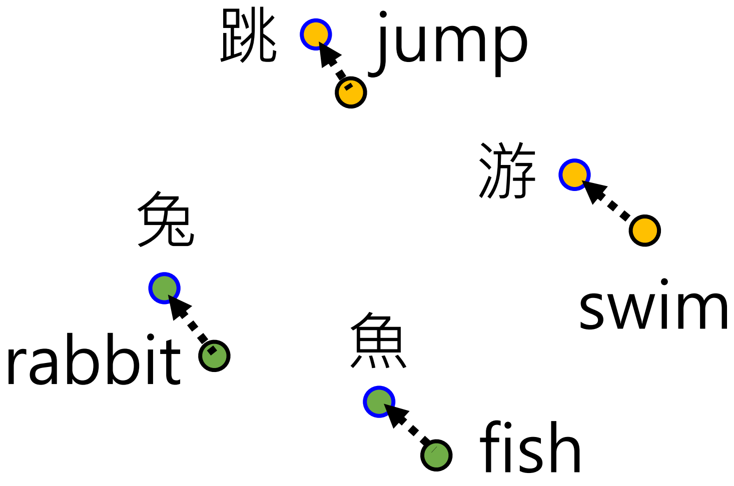
$$z_{long} = \frac{1}{N_1}\sum_{x \in long} En(x) - \frac{1}{N_2}\sum_{x' \notin long} En(x')$$

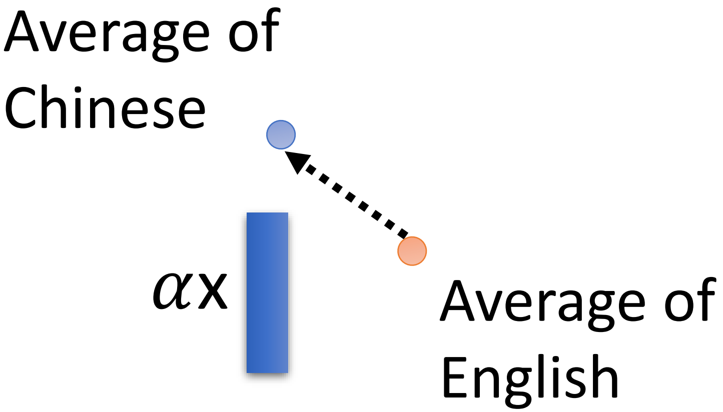
Short Hair $x \rightarrow En(x) + z_{long} = z' \rightarrow Gen(z')$ Long Hair

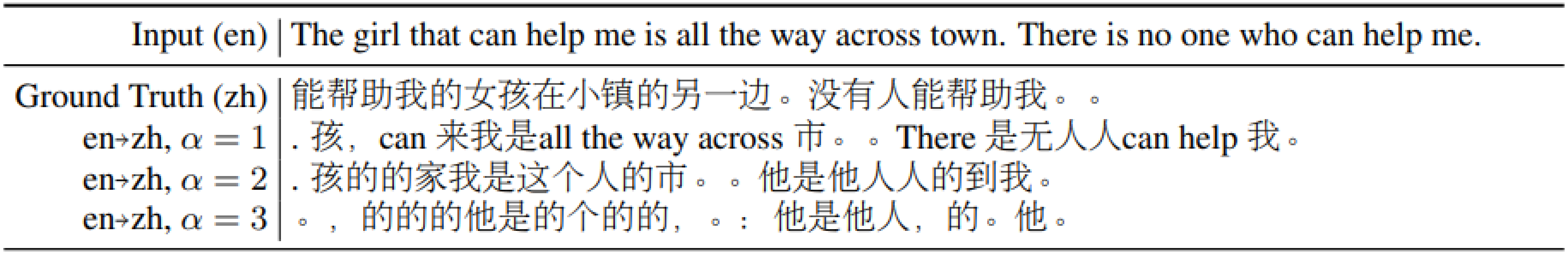
Reference: https://youtu.be/Lhs_Kphd0jg

This work is done by 劉記良、許宗嫄、莊永松

It works!!!

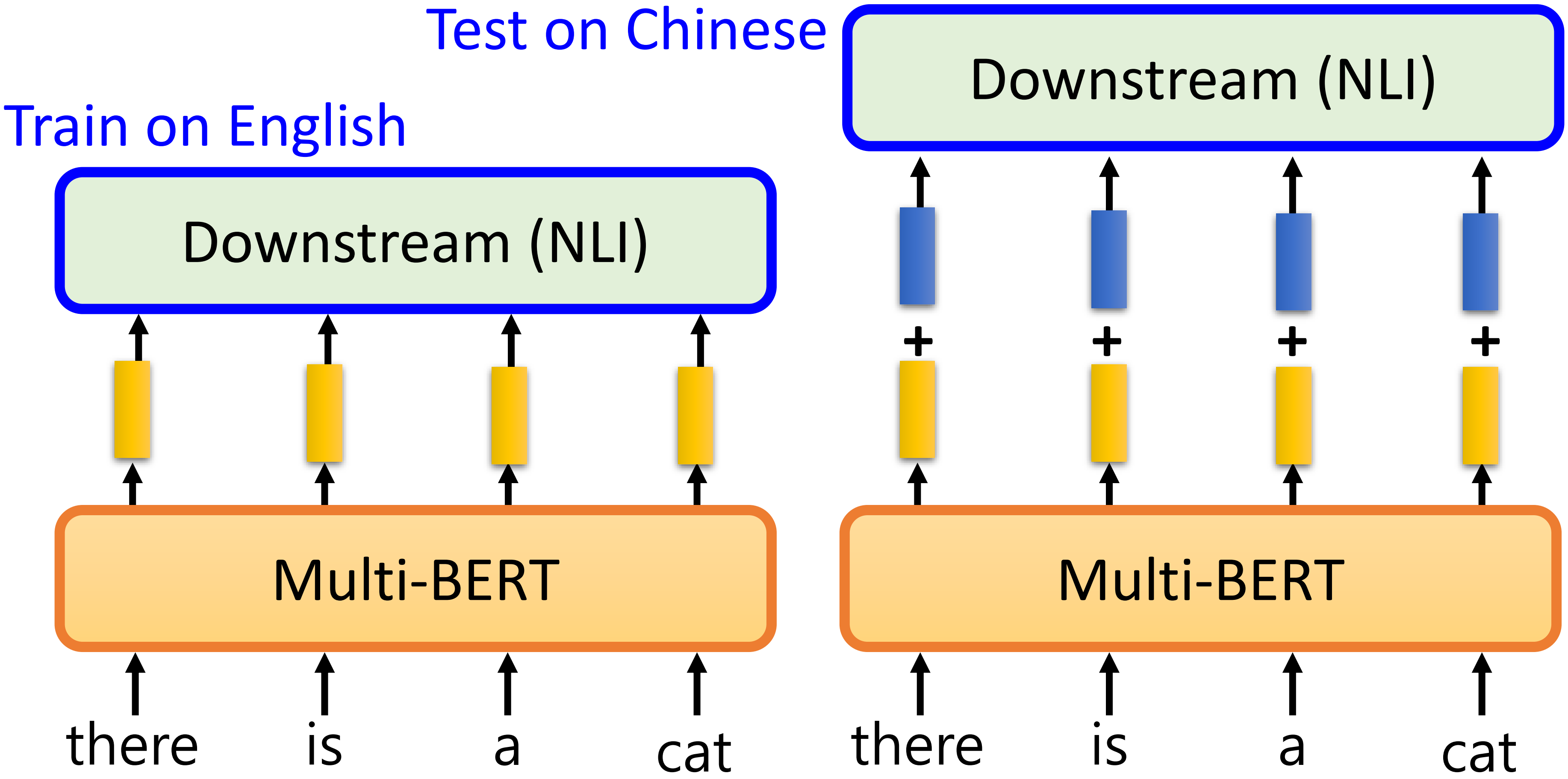






Input (en)	The girl that can help me is all the way across town. There is no one who can help me.
Ground Truth (zh)	能帮助我的女孩在小镇的另一边。没有人能帮助我。。
en→zh, $\alpha = 1$	. 孩，can 来我是all the way across 市。。There 是无人人can help 我。
en→zh, $\alpha = 2$	. 孩的的家我是这个人的市。。他是他人人的到我。
en→zh, $\alpha = 3$	。，的的的他是的个的的，。：他是他人，的。他。

[Liu, et al., arXiv’20]



Test on Chinese

Train on English

Downstream (NLI)

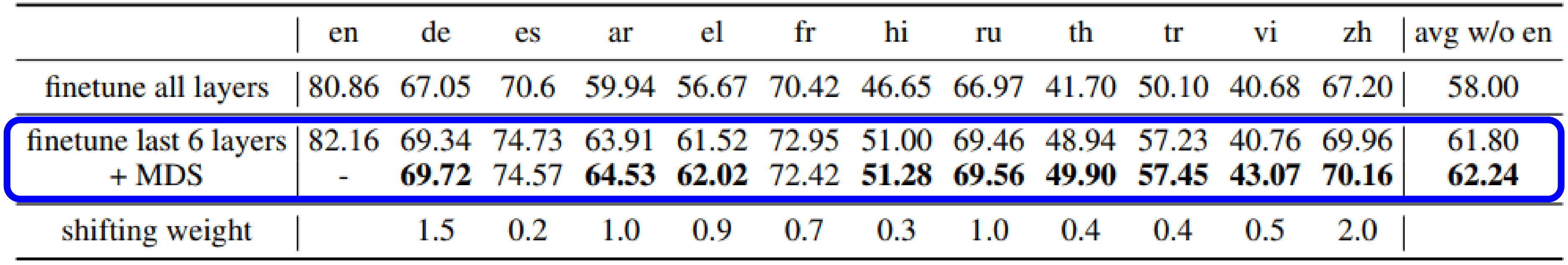
Downstream (NLI)

Multi-BERT

Multi-BERT

there is a cat

there is a cat



	en	de	es	ar	el	fr	hi	ru	th	tr	vi	zh	avg w/o en
finetune all layers	80.86	67.05	70.6	59.94	56.67	70.42	46.65	66.97	41.70	50.10	40.68	67.20	58.00
finetune last 6 layers	82.16	69.34	74.73	63.91	61.52	72.95	51.00	69.46	48.94	57.23	40.76	69.96	61.80
+ MDS	-	**69.72**	74.57	**64.53**	**62.02**	72.42	**51.28**	**69.56**	**49.90**	**57.45**	**43.07**	**70.16**	**62.24**
shifting weight		1.5	0.2	1.0	0.9	0.7	0.3	1.0	0.4	0.4	0.5	2.0	

Reference

- [K, et al., ICLR’20] Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual BERT: An empirical study, ICLR, 2020
- [Pires, et al., ACL’19] Telmo Pires, Eva Schlinger, Dan Garrette, How multilingual is Multilingual BERT?, ACL, 2019
- [Wu, et al., EMNLP’19] Shijie Wu, Mark Dredze, Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT, EMNLP, 2019
- [Hsu, Liu, et al., EMNLP’19] Tsung-Yuan Hsu, Chi-Liang Liu and Hung-yi Lee, "Zero-shot Reading Comprehension by Cross-lingual Transfer Learning with Multi-lingual Language Representation Model", EMNLP, 2019
- [Liu, et al., arXiv’20] Chi-Liang Liu, Tsung-Yuan Hsu, Yung-Sung Chuang, Hung-Yi Lee, A Study of Cross-Lingual Ability and Language-specific Information in Multilingual BERT, arXiv, 2020

Reference

- [Hu, et al., arXiv’20] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, Melvin Johnson, XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization, arXiv, 2020
- [Libovický, arXiv’20] Jindřich Libovický, Rudolf Rosa, Alexander Fraser, On the Language Neutrality of Pre-trained Multilingual Representations, arXiv, 2020