A few words about BERT

(or what’s up with NLP field)

based on the original article https://arxiv.org/abs/1810.04805

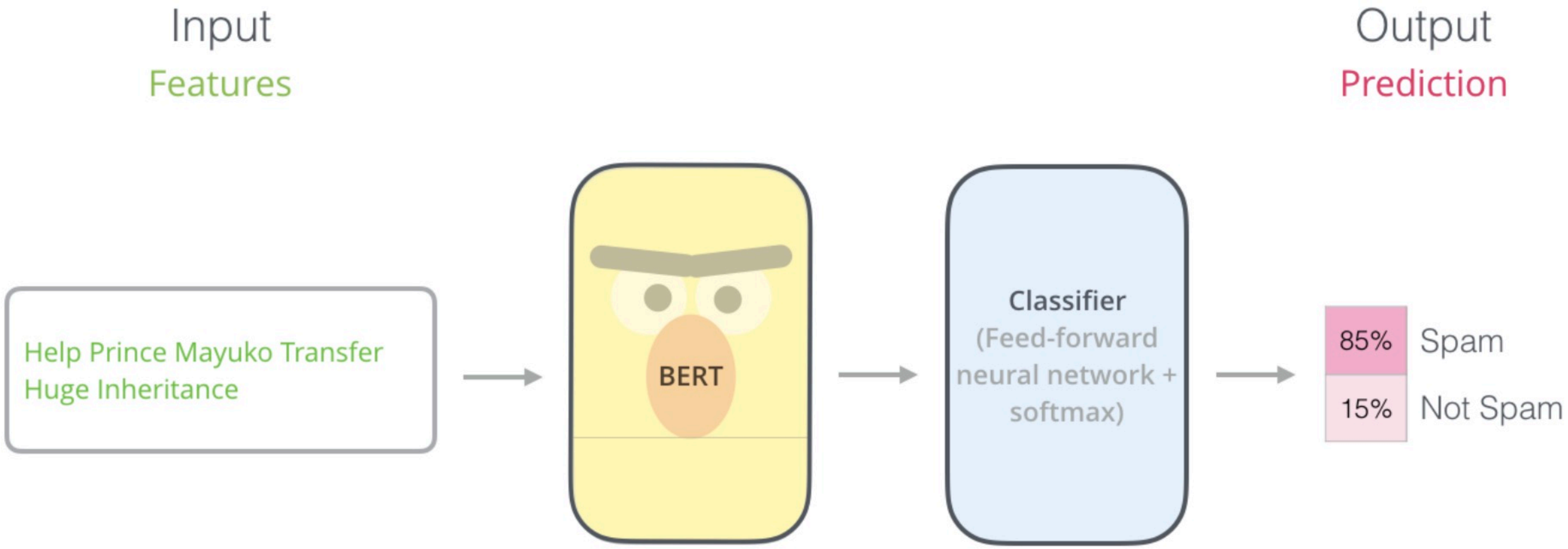
Sochi, 2019

Outline

- BERT Model
- Benchmarks for BERT
- Details of BERT Model
- Training framework
- BERT vs XLNet
- RoBERTa by FAIR

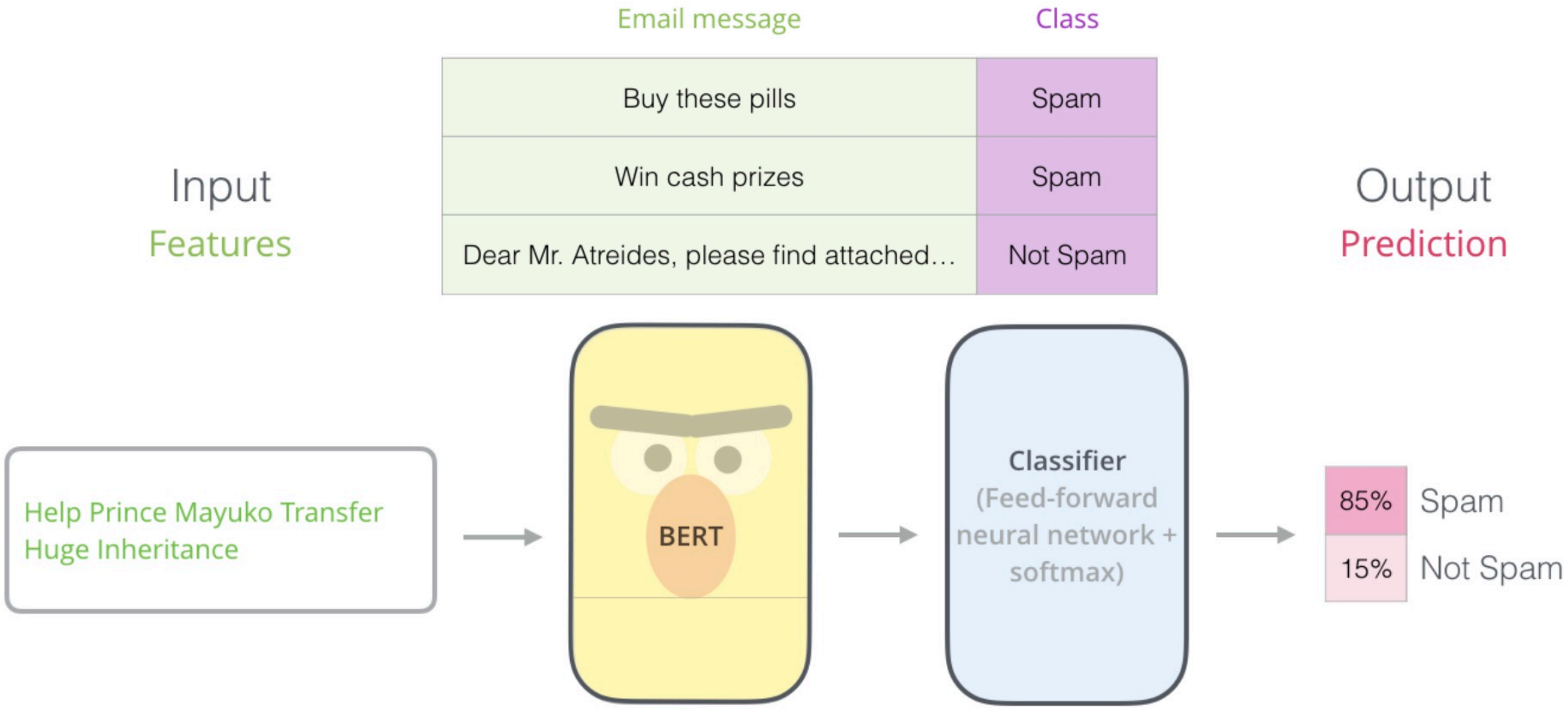
https://arxiv.org/abs/1810.04805

BERT Model


Input
Features
Help Prince Mayuko Transfer Huge Inheritance
BERT
Classifier
(Feed-forward neural network + softmax)
Output
Prediction
85%
Spam
15%
Not Spam

Source: http://jalammar.github.io/illustrated-bert/

BERT Model

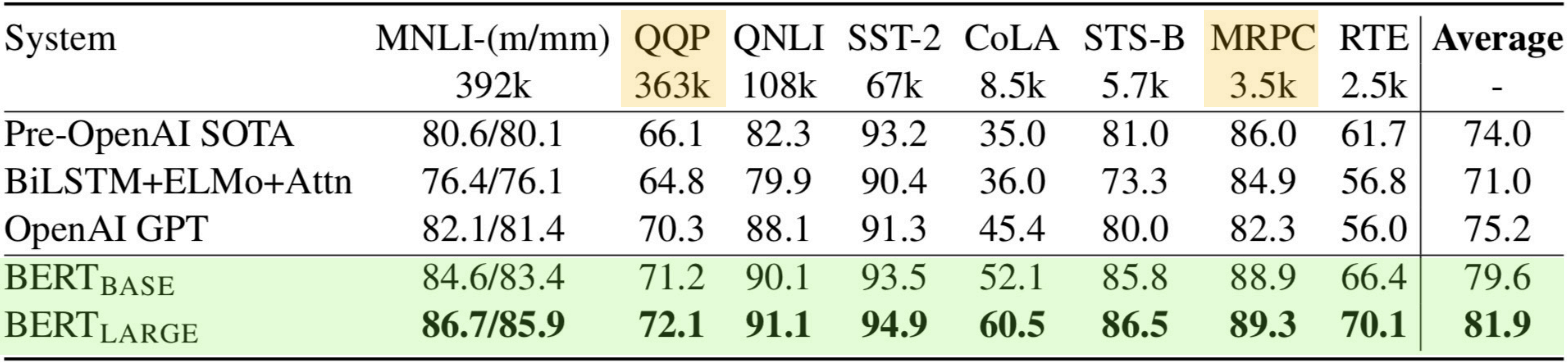


Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached…	Not Spam

Input
Features
Output
Prediction
Help Prince Mayuko Transfer Huge Inheritance
BERT
Classifier
(Feed-forward neural network + softmax)
85% Spam
15% Not Spam

Source: http://jalammar.github.io/illustrated-bert/

Benchmarks for BERT



System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
BERT$_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
BERT$_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

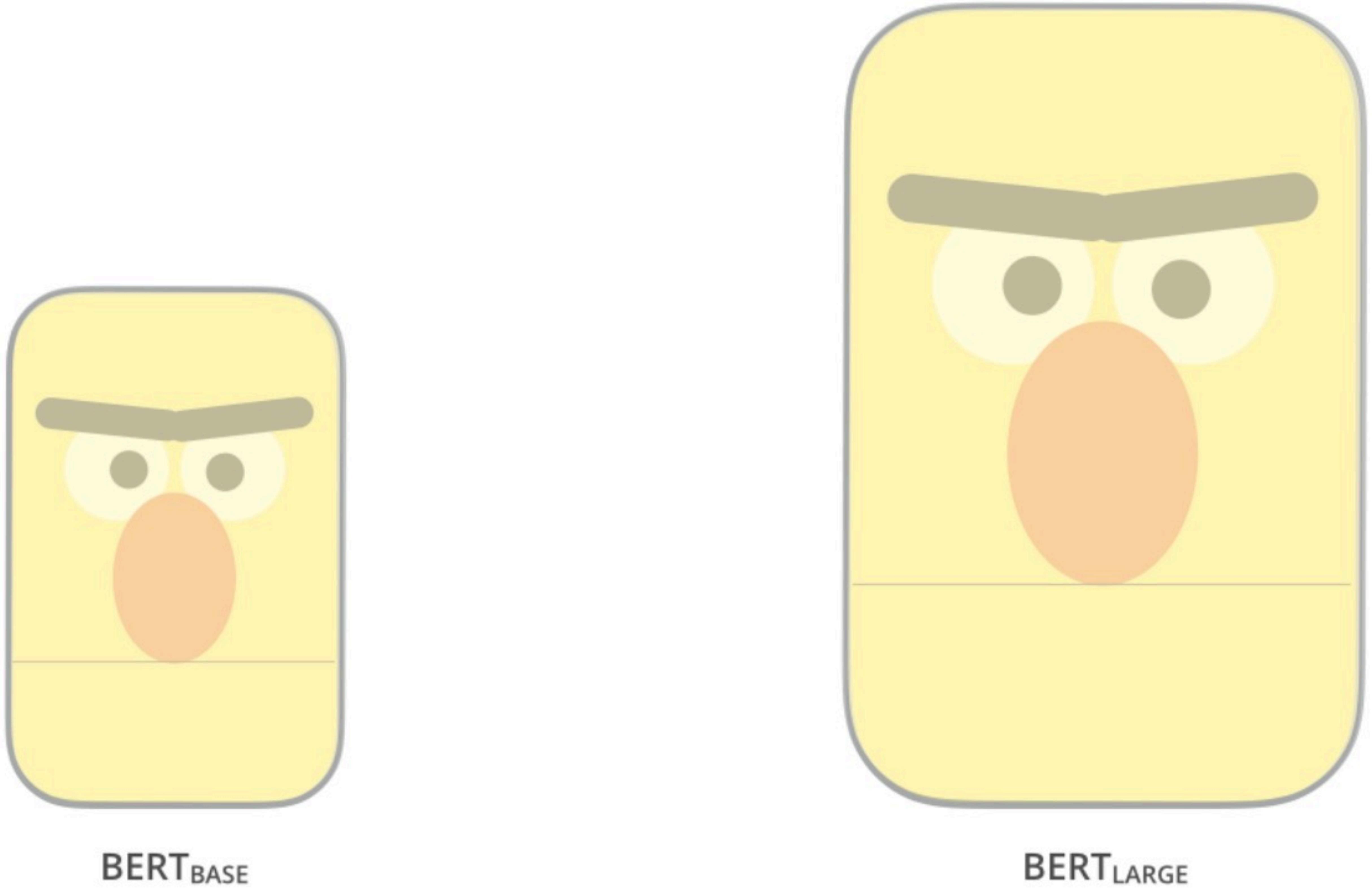
- **QQP** - Quora Question Pairs

- **MRPC** - Microsoft Research Paraphrase Corpus

- …

https://arxiv.org/abs/1810.04805

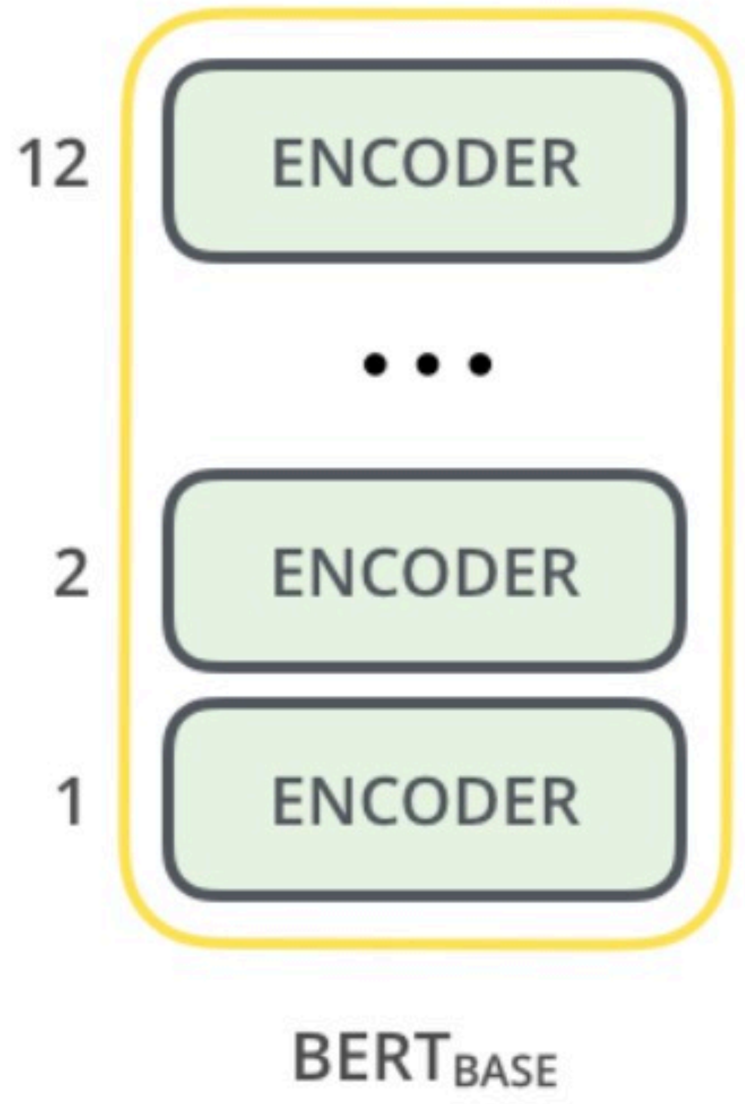
Details of BERT Model



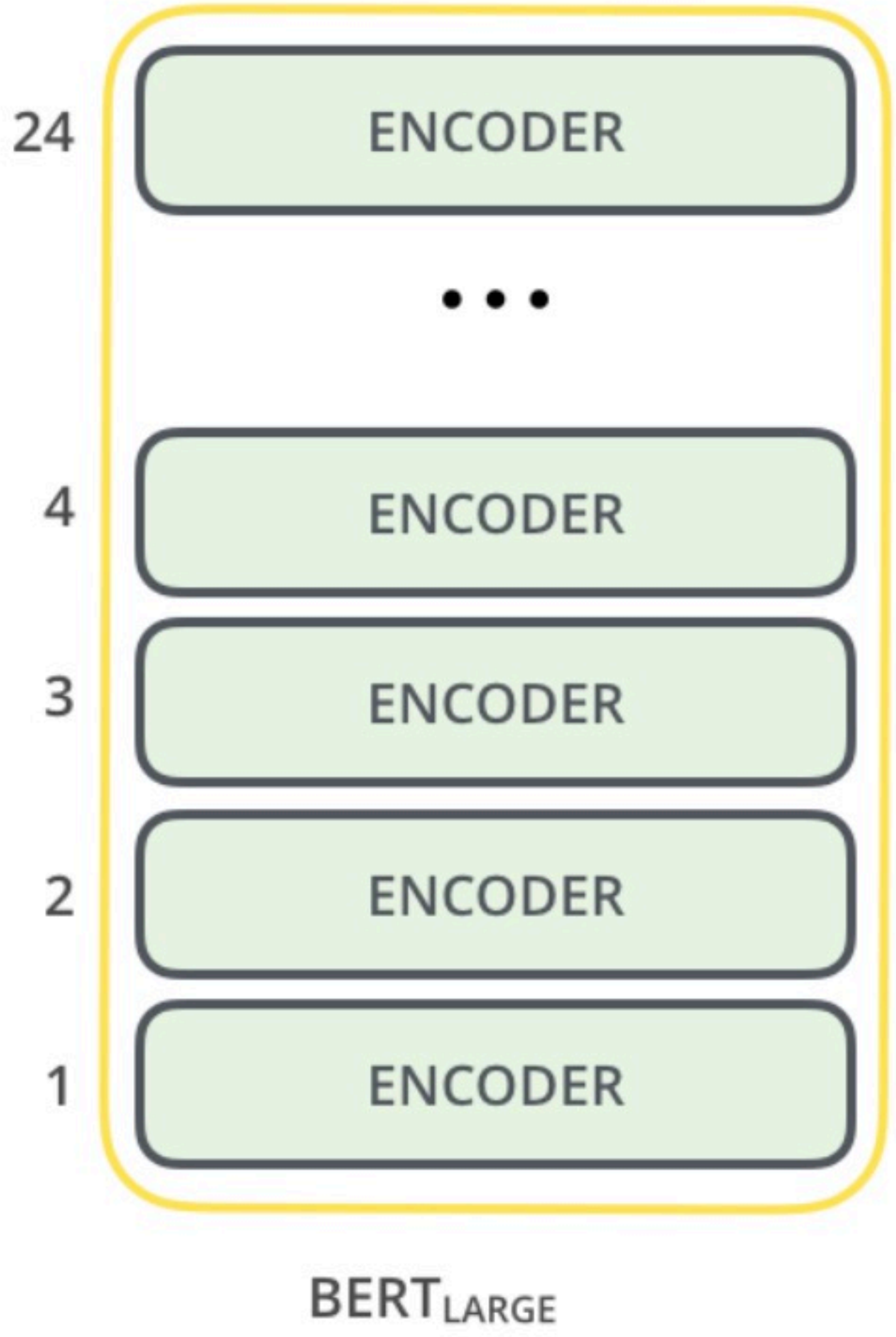
Source: http://jalammar.github.io/illustrated-bert/

Details of BERT Model

With total parameters of **~110M**

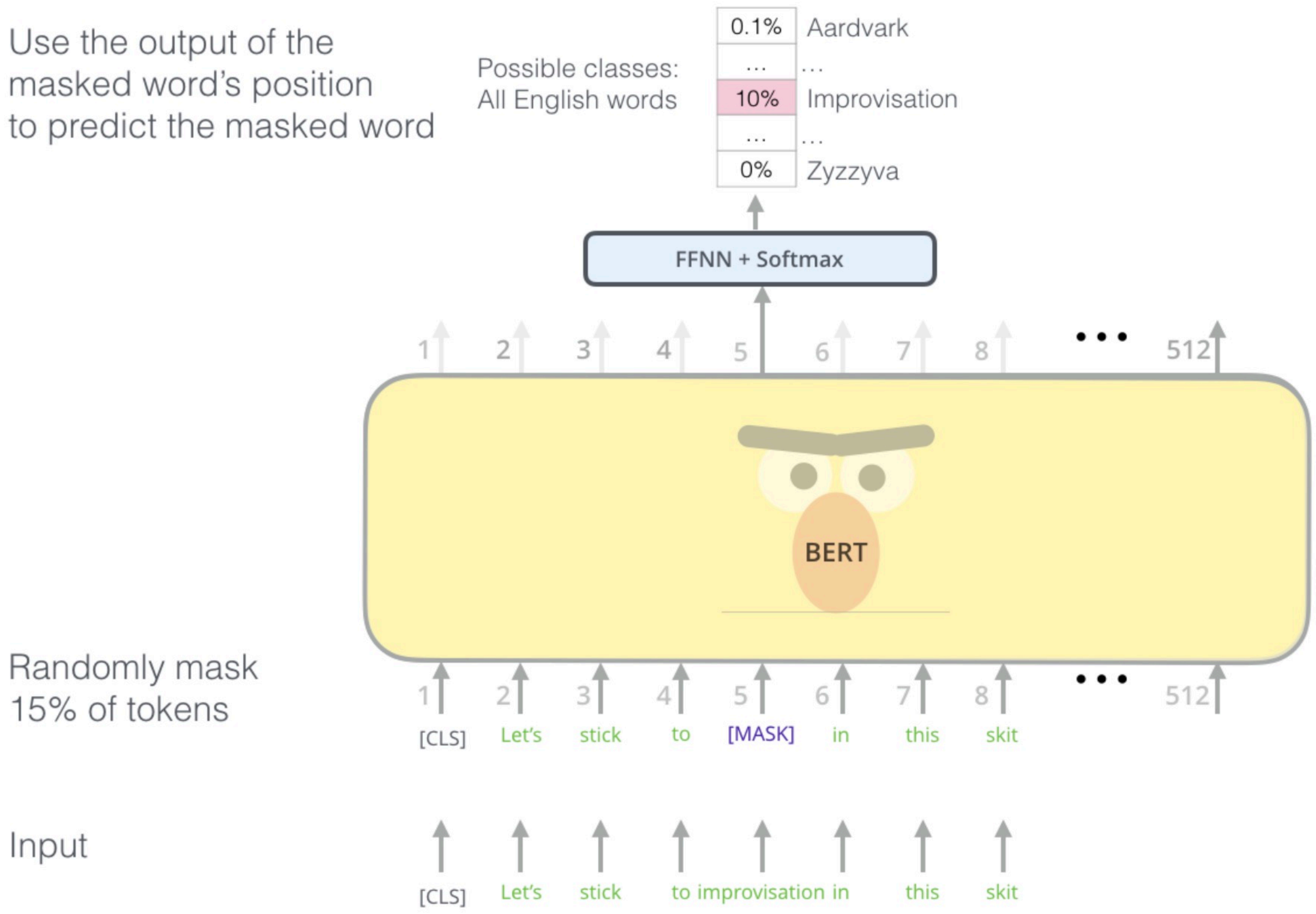


With total parameters of **~340M**



Source: http://jalammar.github.io/illustrated-bert/

Details of BERT Model



Use the output of the masked word's position to predict the masked word

Possible classes: All English words

0.1%	Aardvark
…	…
10%	Improvisation
…	…
0%	Zyzzyva

FFNN + Softmax

Randomly mask 15% of tokens

[CLS] Let's stick to [MASK] in this skit

Input

[CLS] Let's stick to improvisation in this skit

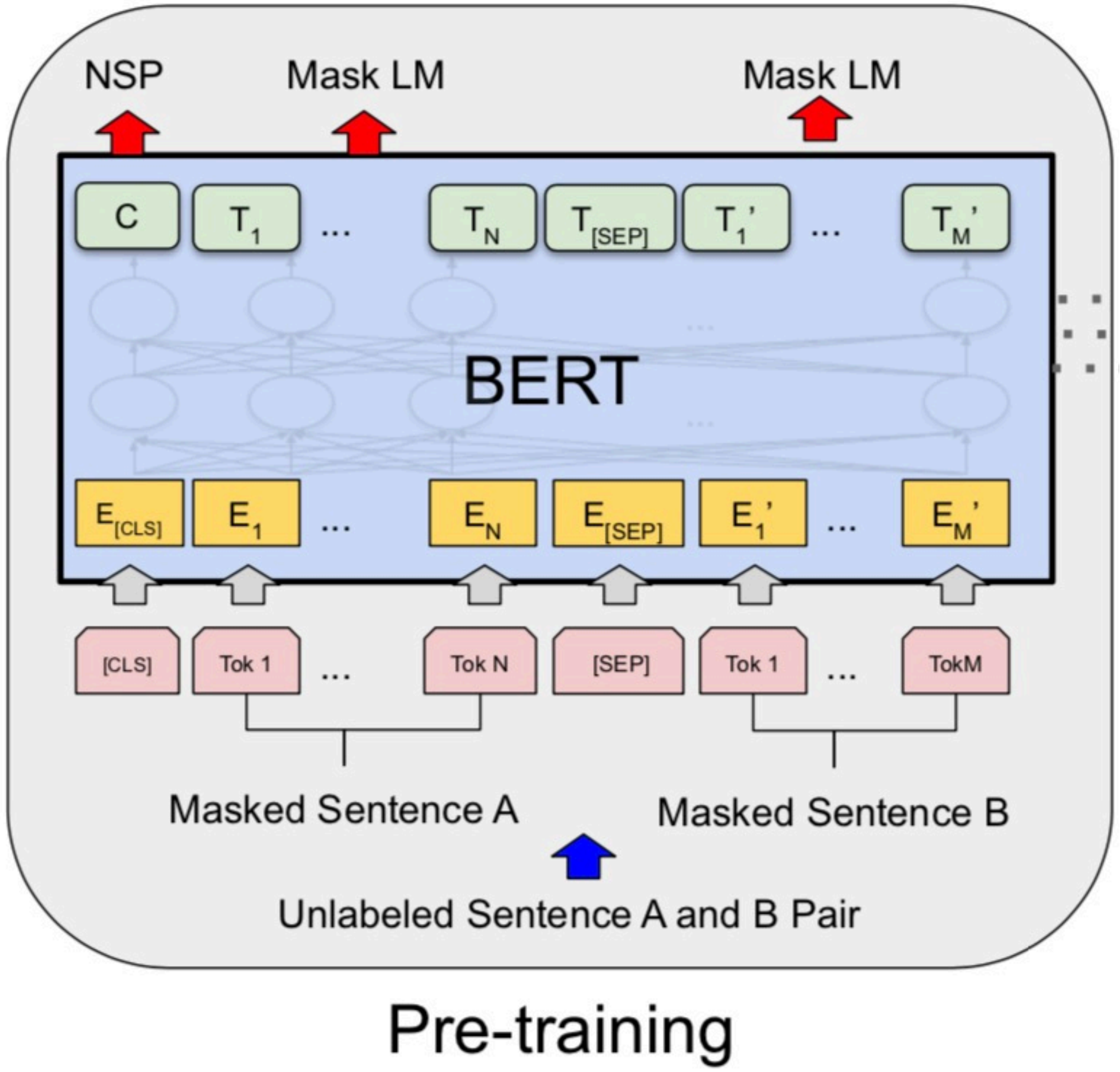
Source: http://jalammar.github.io/illustrated-bert/

Training framework

- **Pre-training.** Model is trained on unlabelled data over different pre-training tasks.
- **Fine-tuning.** Firstly, model is initialised with the pre-trained parameters. After that all the parameters are fine-tuned using labelled data from the downstream tasks.

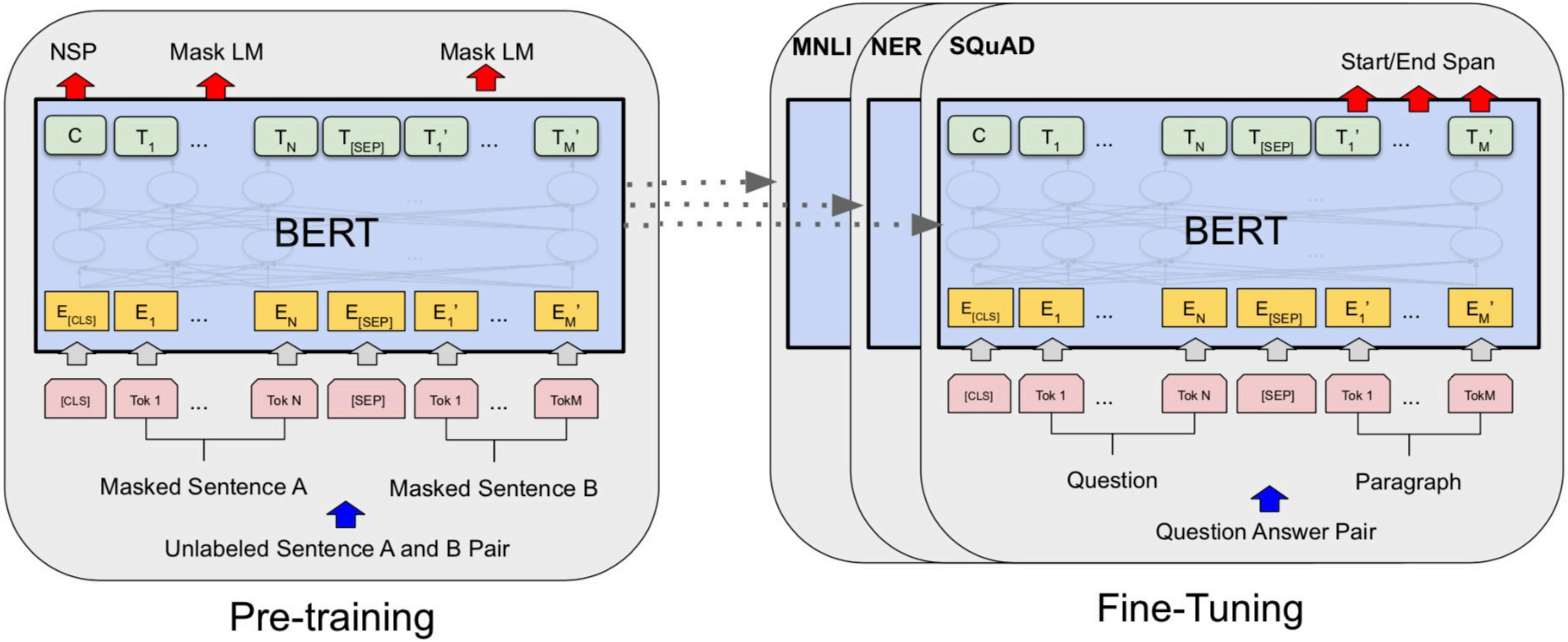
https://arxiv.org/abs/1810.04805

Training framework



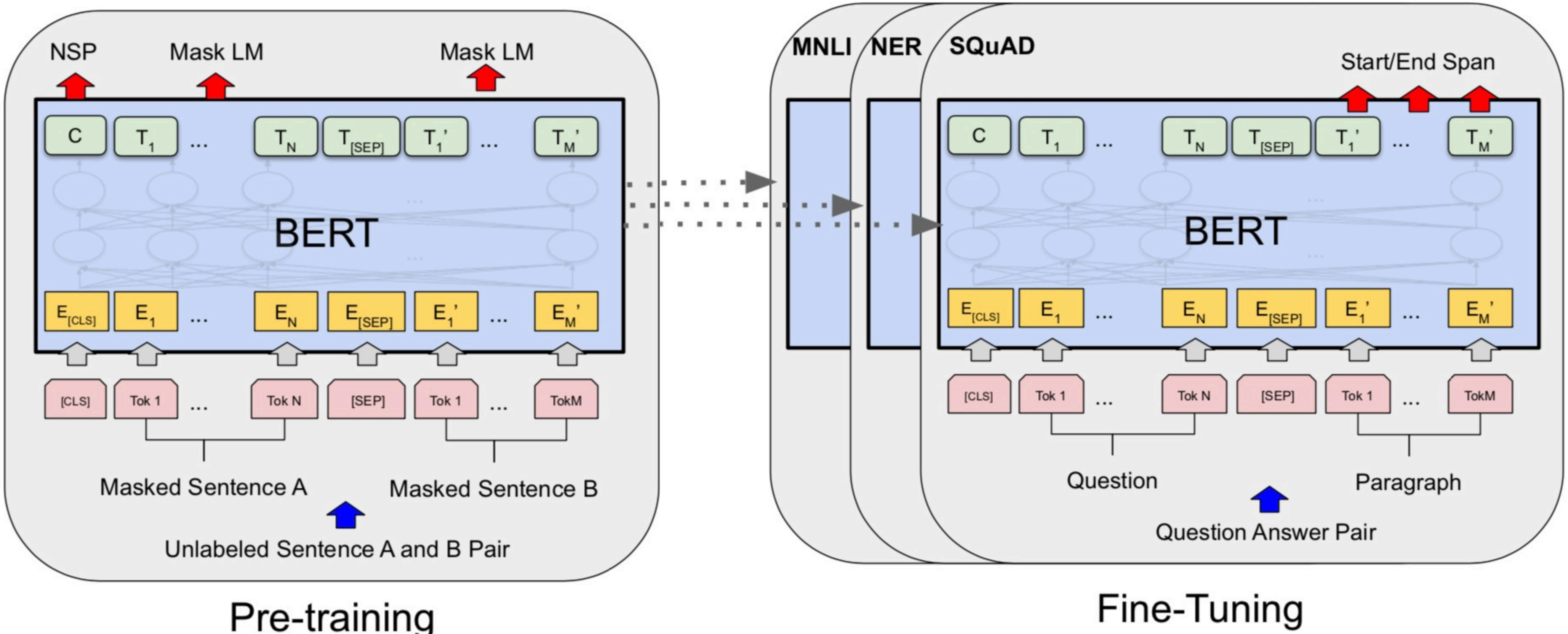
https://arxiv.org/abs/1810.04805

Training framework



https://arxiv.org/abs/1810.04805

Training framework

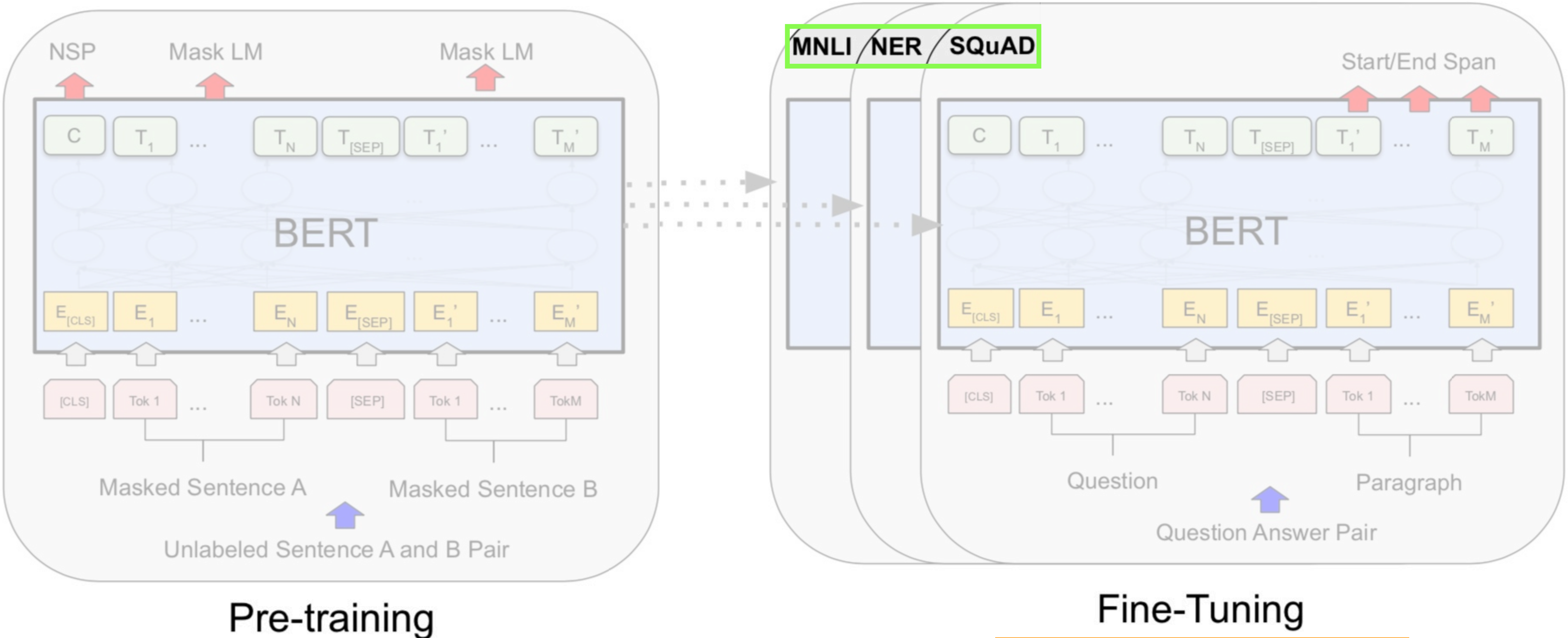


Pair of sentences **A** and **B**

Pair of **Q**uestion and **A**nswer

https://arxiv.org/abs/1810.04805

Training framework



Pair of **Q**uestion and **A**nswer

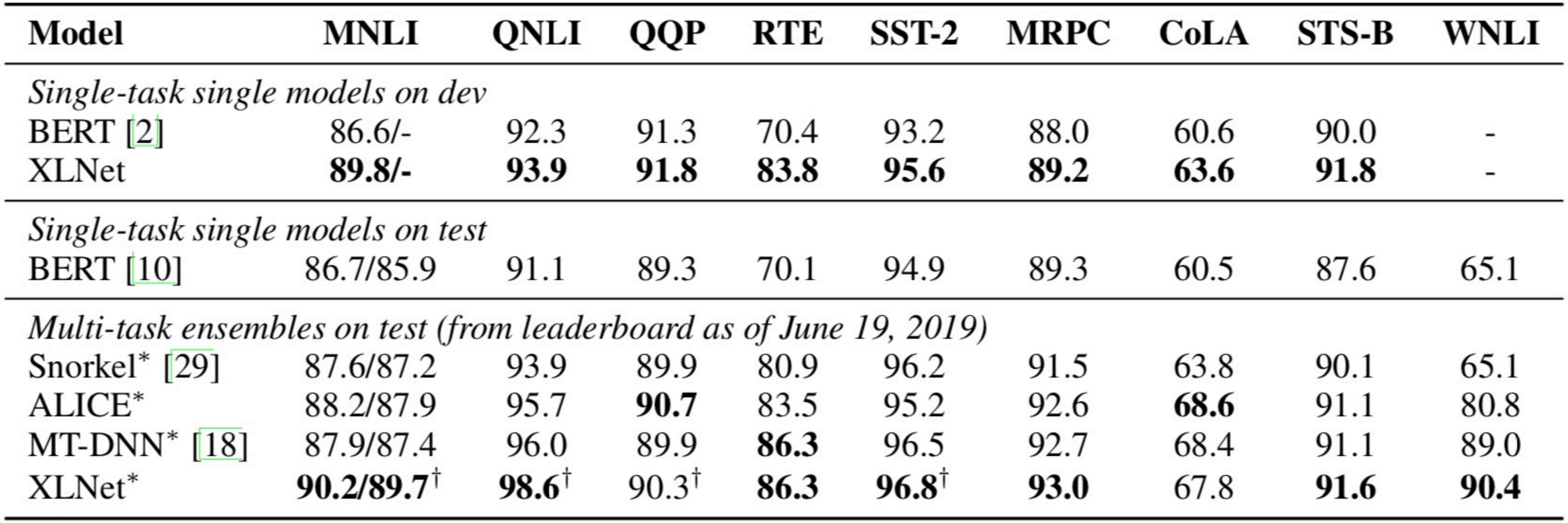
https://arxiv.org/abs/1810.04805

BERT vs XLNet

- BERT - **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (looses connection between words)

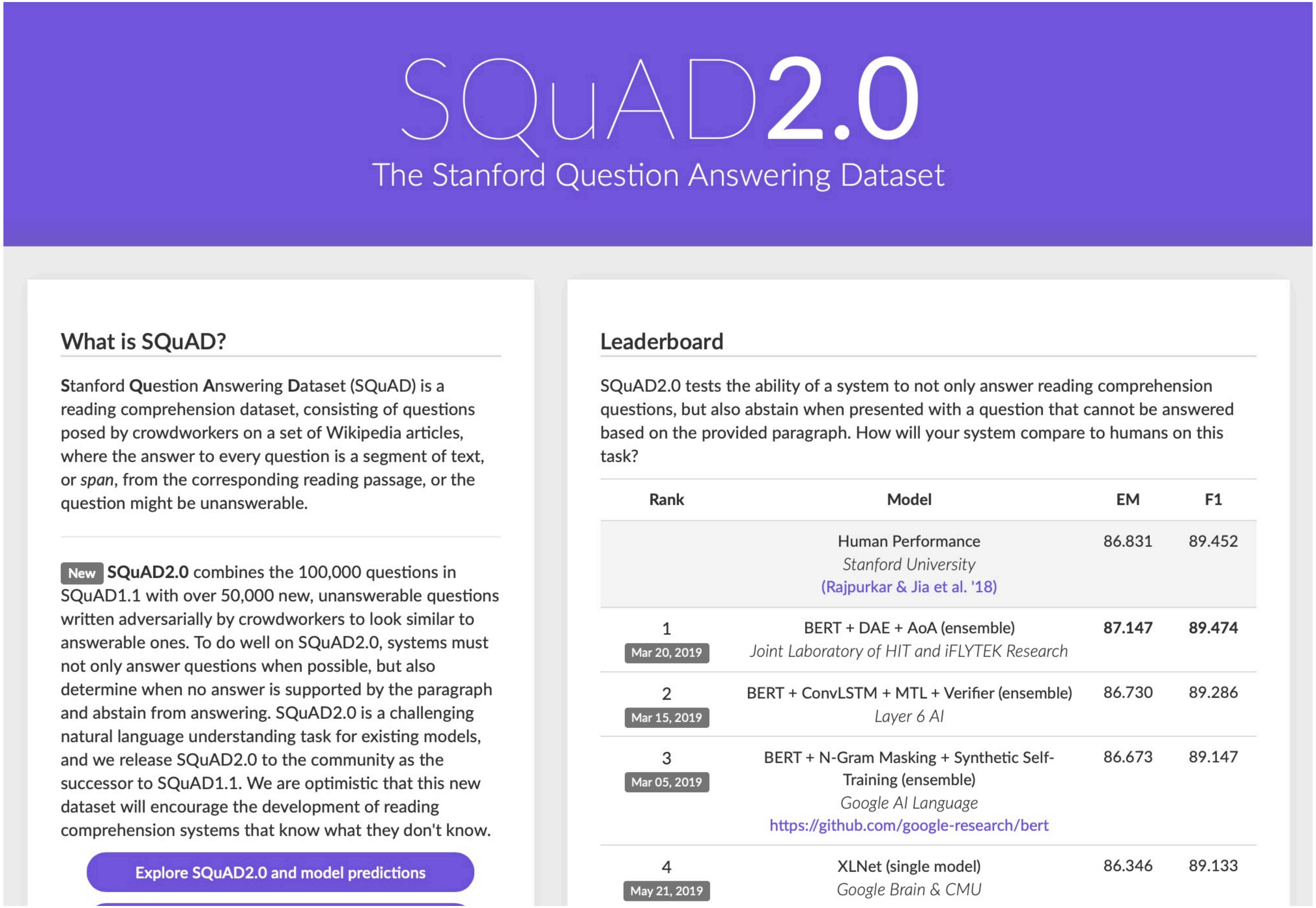
- XLNet - AutoRegressive Language Modelling (word order does count)

BERT vs XLNet



Model	MNLI	QNLI	QQP	RTE	SST-2	MRPC	CoLA	STS-B	WNLI
Single-task single models on dev									
BERT [2]	86.6/-	92.3	91.3	70.4	93.2	88.0	60.6	90.0	-
XLNet	**89.8/-**	**93.9**	**91.8**	**83.8**	**95.6**	**89.2**	**63.6**	**91.8**	-
Single-task single models on test									
BERT [10]	86.7/85.9	91.1	89.3	70.1	94.9	89.3	60.5	87.6	65.1
Multi-task ensembles on test (from leaderboard as of June 19, 2019)									
Snorkel* [29]	87.6/87.2	93.9	89.9	80.9	96.2	91.5	63.8	90.1	65.1
ALICE*	88.2/87.9	95.7	**90.7**	83.5	95.2	92.6	**68.6**	91.1	80.8
MT-DNN* [18]	87.9/87.4	96.0	89.9	**86.3**	96.5	92.7	68.4	91.1	89.0
XLNet*	**90.2/89.7**†	**98.6**†	90.3†	**86.3**	**96.8**†	**93.0**	67.8	**91.6**	**90.4**

Source: https://arxiv.org/abs/1906.08237



SQuAD2.0

The Stanford Question Answering Dataset

What is SQuAD?

Stanford **Qu**estion **A**nswering **D**ataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, or the question might be unanswerable.

New **SQuAD2.0** combines the 100,000 questions in SQuAD1.1 with over 50,000 new, unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering. SQuAD2.0 is a challenging natural language understanding task for existing models, and we release SQuAD2.0 to the community as the successor to SQuAD1.1. We are optimistic that this new dataset will encourage the development of reading comprehension systems that know what they don't know.

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph. How will your system compare to humans on this task?

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	**87.147**	**89.474**
2 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
3 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	86.673	89.147
4 May 21, 2019	XLNet (single model) *Google Brain & CMU*	86.346	89.133

Source: https://rajpurkar.github.io/SQuAD-explorer/

RoBERTa by FAIR

GLUE
SuperGLUE
Tasks
Leaderboard
FAQ
Diagnostics
Submit
Login

	Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
	1	Facebook AI	RoBERTa		88.5	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	98.9	88.2	89.0	48.7
	2	XLNet Team	XLNet-Large (ensemble)	↗	88.4	67.8	96.8	93.0/90.7	91.6/91.1	74.2/90.3	90.2	89.8	98.6	86.3	90.4	47.5
+	3	Microsoft D365 AI & MSR AI	MT-DNN-ensemble	↗	87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0	42.8
	4	GLUE Human Baselines	GLUE Human Baselines	↗	87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9	-
+	5	王玮	ALICE large ensemble (Alibaba DAMO NL		86.3	68.6	95.2	92.6/90.2	91.1/90.6	74.4/90.7	88.2	87.9	95.7	83.5	80.8	43.9

Facebook AI **July 19, 2019**

Like This Page · 7 hrs ·

Congratulations to our AI team for matching the top GLUE benchmark performance! We believe strongly in open & collaborative research and thank @GoogleAI for releasing BERT. It led to RoBERTa, our robustly optimized system that was trained longer, on more data.
https://gluebenchmark.com/leaderboard/

Daniel S. ReMine Nice! Thanks for sharing FAIR. Now we need Ro-XLNet-a. 😉

Kevin Wood Yes, keep yup the

Summary

- BERT Model
- Benchmarks for BERT
- Details of BERT Model
- Training framework
- BERT vs XLNet
- RoBERTa by FAIR

References

- The Illustrated BERT, ELMo, and co.: http://jalammar.github.io/illustrated-bert/
- About BERT in Google AI Blog: https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html
- SQuAD 2.0: https://rajpurkar.github.io/SQuAD-explorer/
- GLUE Benchmark: https://gluebenchmark.com/leaderboard