Natural Language Processing (9)

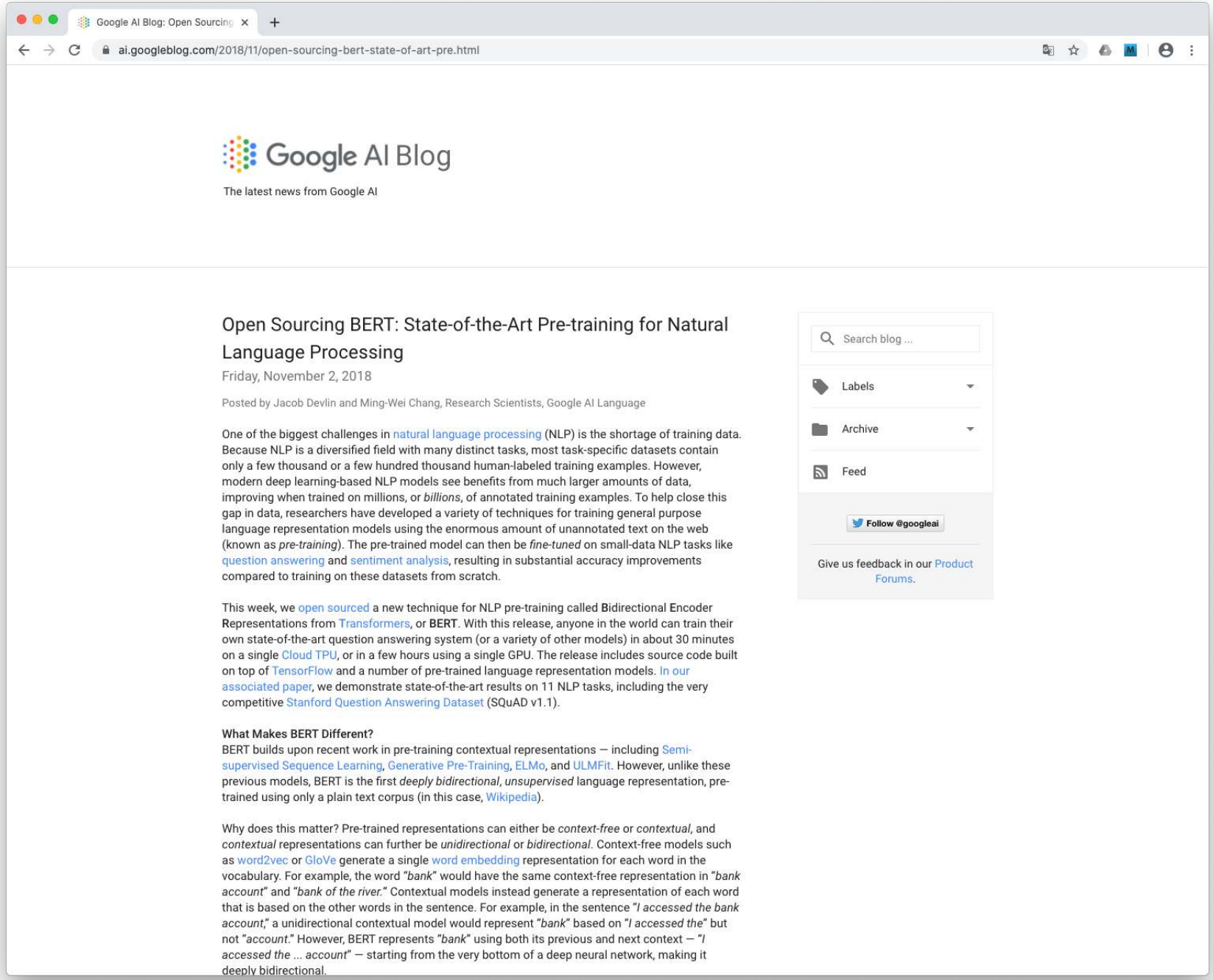
Transfer Learning

Daisuke Kawahara

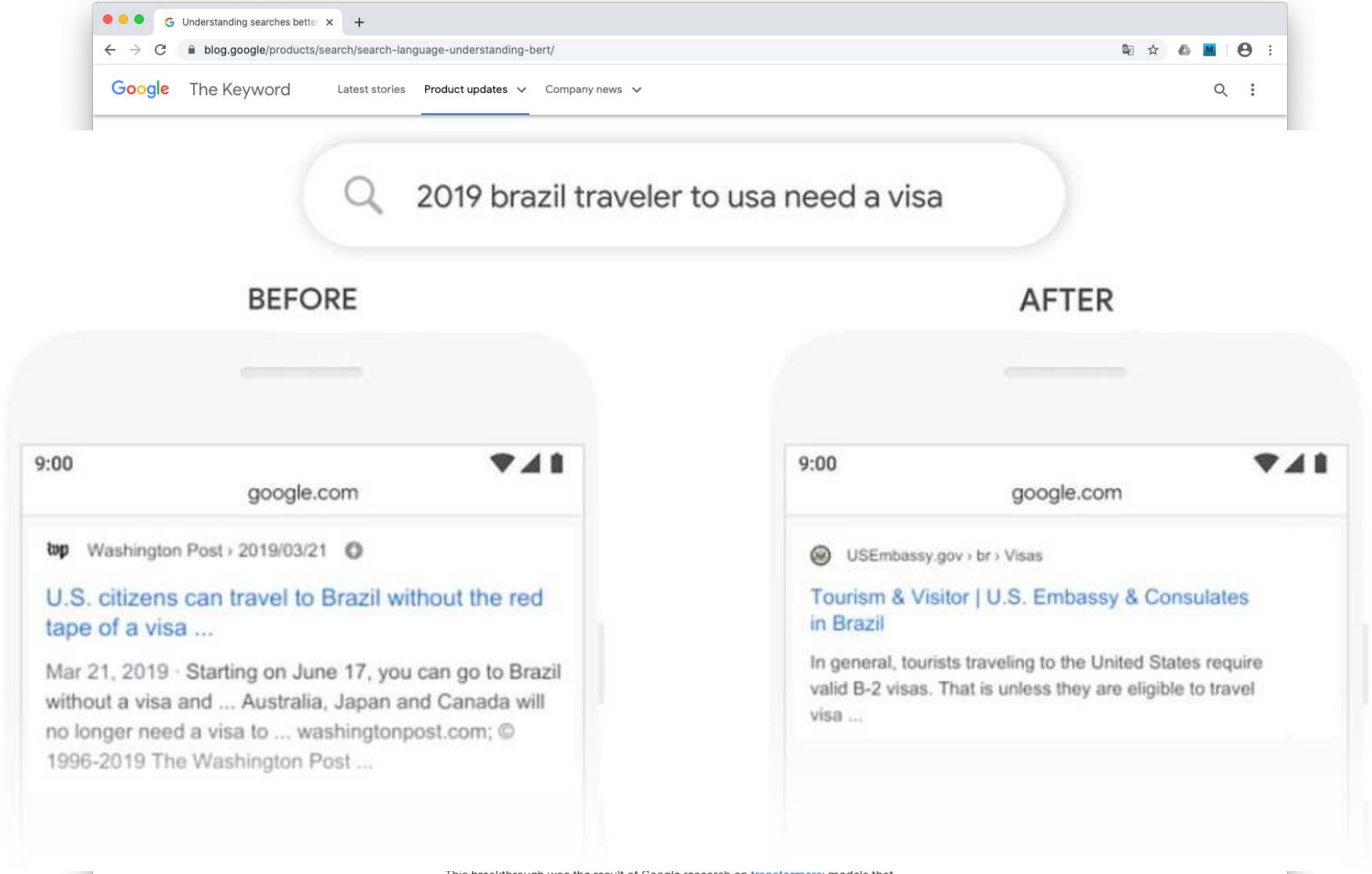
Department of Communications and Computer Engineering,
Waseda University

Lecture Plan

1. Overview of Natural Language Processing
2. Formal Language Theory
3. Word Senses and Embeddings
4. Topic Models
5. Collocations, Language Models, and Recurrent Neural Networks
6. Sequence Labeling and Morphological Analysis
7. Parsing (1): Constituency Parsing
8. Parsing (2): Dependency Parsing
9. Transfer Learning
10. Knowledge Acquisition
11. Information Retrieval, Question Answering, and Machine Translation
12. Guest Talk (1)
13. Guest Talk (2)
14. Project: Survey or Programming
15. Project Presentation



https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html



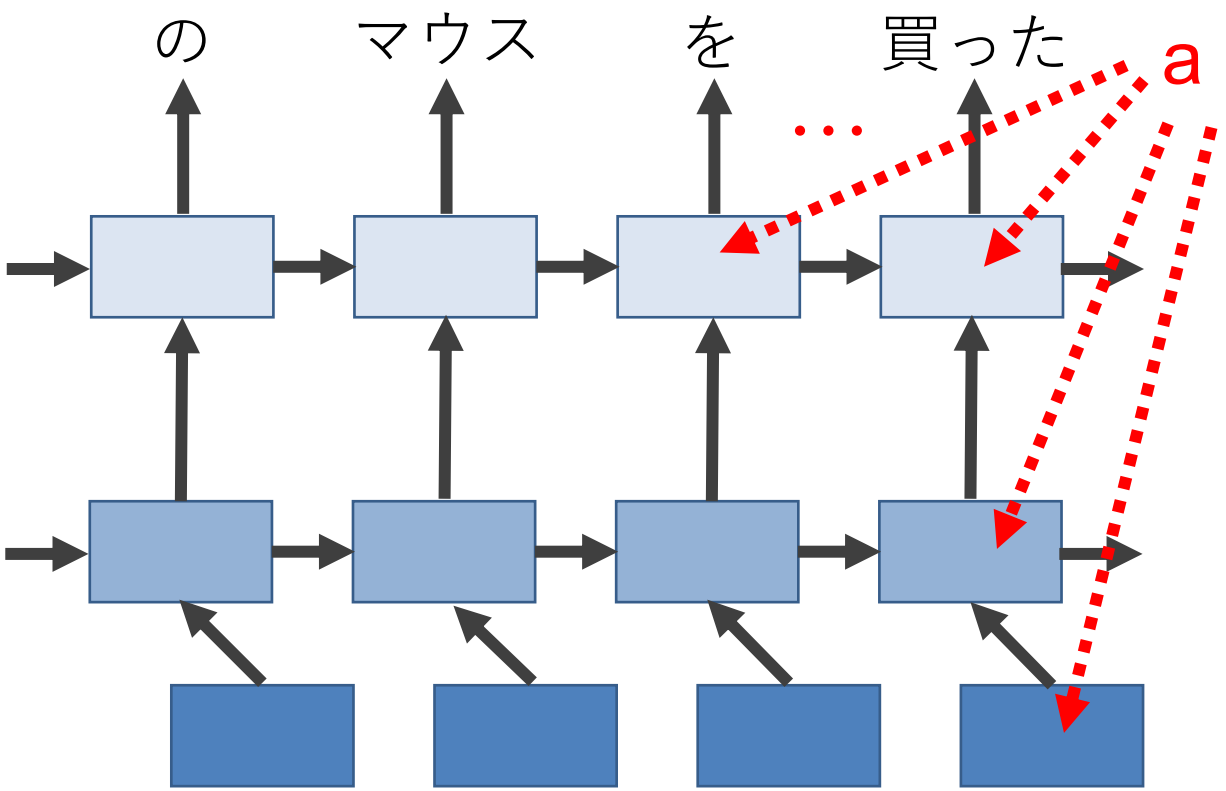
https://www.blog.google/products/search/search-language-understanding-bert/

ELMo [Peters+ 2018]
Embeddings from Language Models

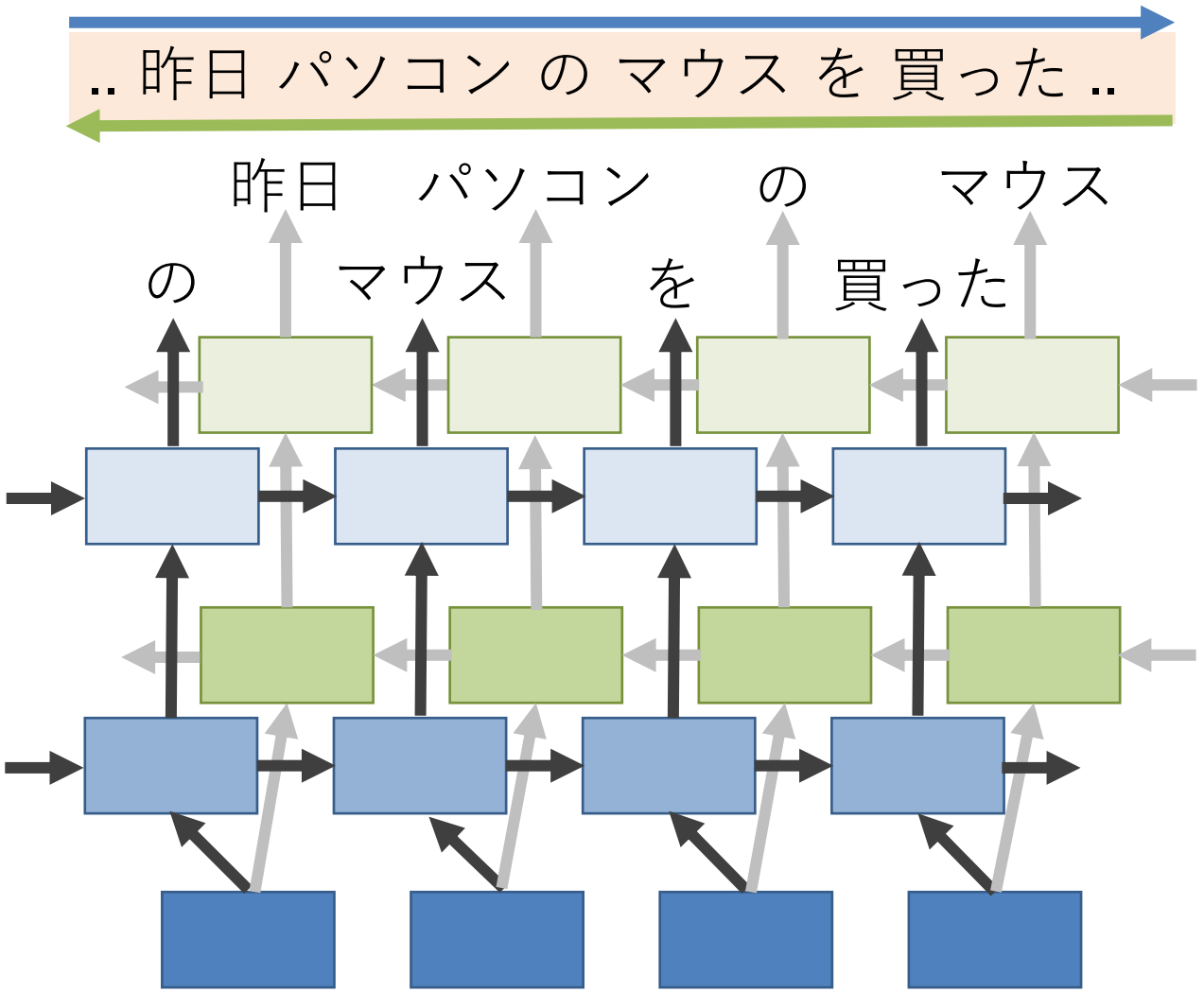
- Purpose
  1. To learn word meanings according to a context
  2. To learn language representations that consider syntax and semantics
- Method
  - Learn a bidirectional language model from a large corpus to obtain language representations (pre-training)
    - Note that language model learning is not a purpose
- Result
  - Achieved SOTA on six tasks

Pre-training: Learn a bidirectional language model

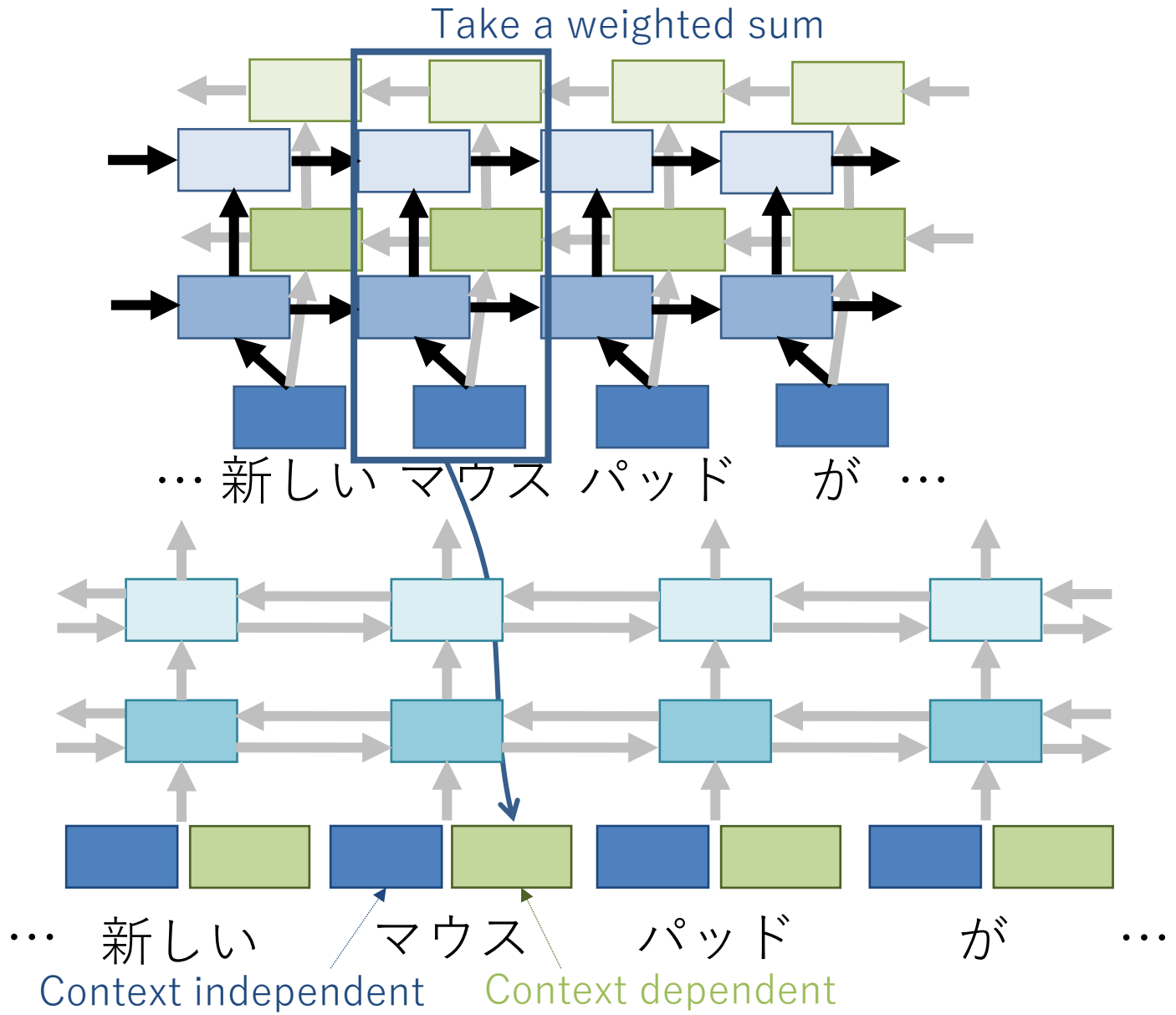
.. 昨日 パソコン の マウス を 買った ..



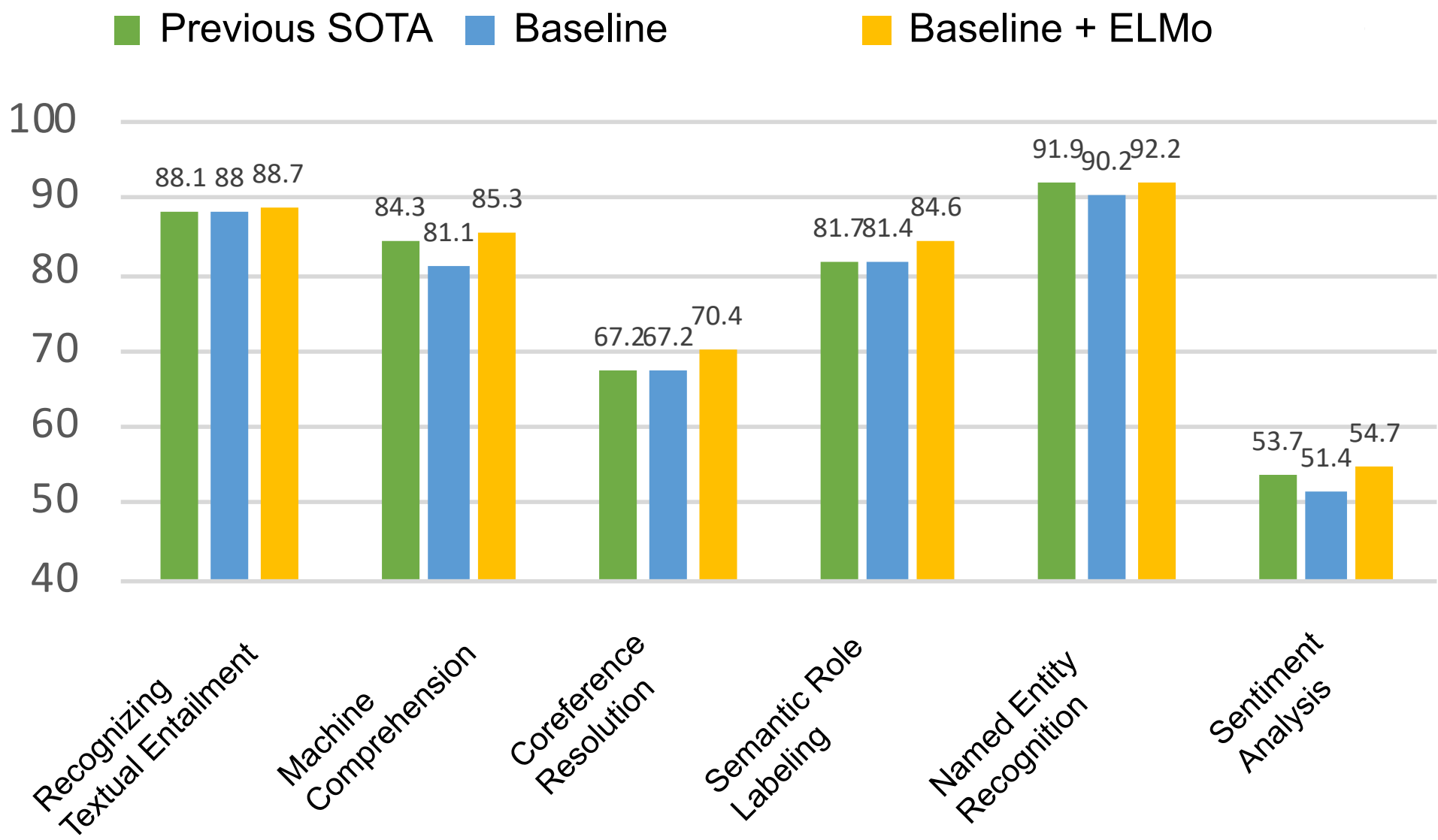
Pre-training: Learn a bidirectional language model



Use in a Downstream Task



Experimental Results



BERT [Devlin+ 2019]

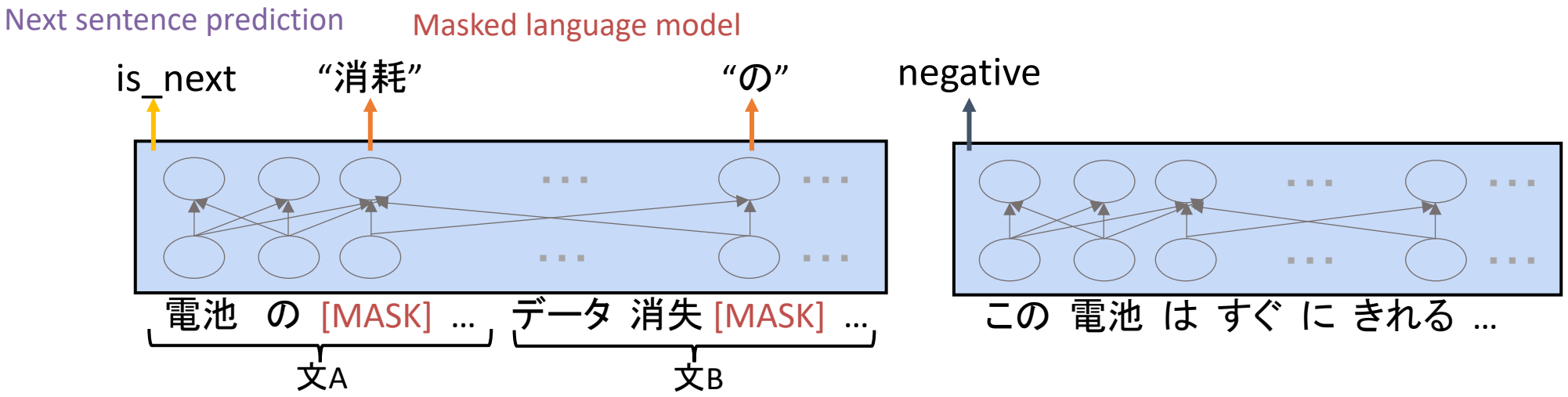
- Model: Transformer [Vaswani+ 2017]
- Training: two steps (pre-training and fine-tuning)

1. Pre-training

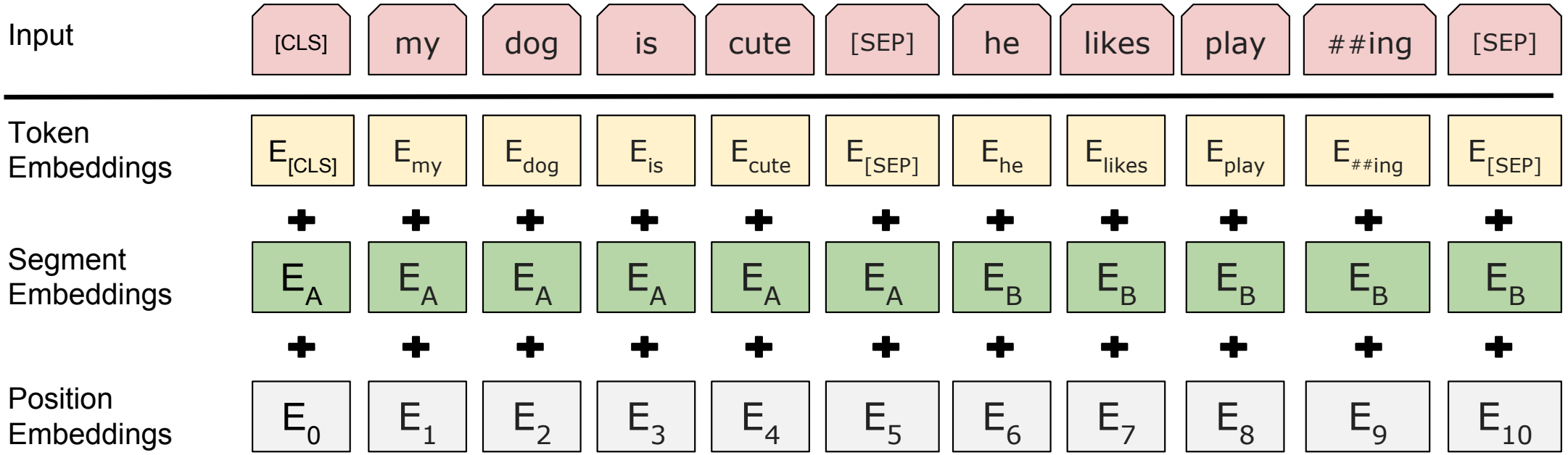
Learn vector representations using a large-scale raw corpus

2. Fine-tuning

Fine-tune parameters for a specific task (e.g., sentiment analysis)



Input Representation



Subwords

- A unit between a word and a character
- The purpose is to reduce unknown words
  - If subwords are not used, words other than the vocabulary (e.g., 30K words) become the [UNK] token
- Subwords have been introduced in neural machine translation

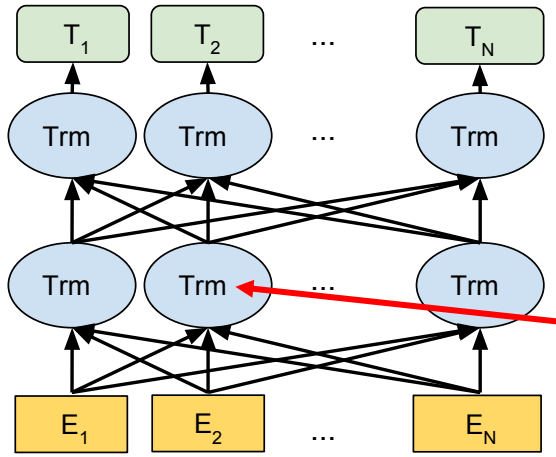
- An algorithm to induce subwords is Byte Pair Encoding (BPE) [Sennrich+ 2016]

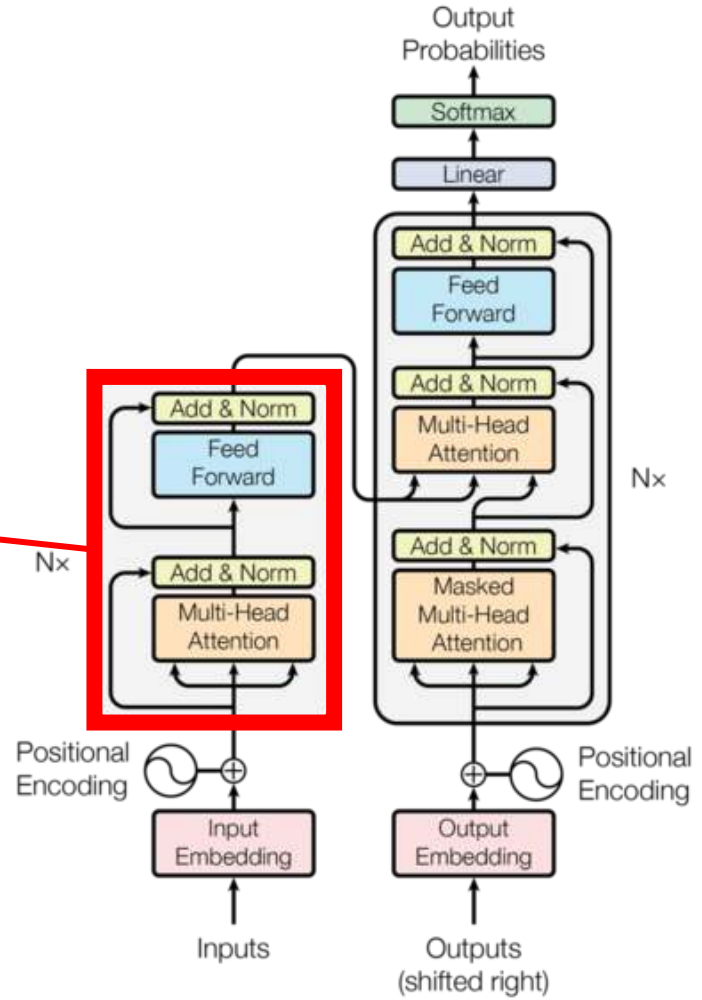
Algorithm of BPE

1. Regard each character as a subword
2. Find a bigram (a pair of subword) with the highest frequency
3. Regard the bigram as a subword
4. Go to Step 2

Freq					
4	l o w	l o w	l o w	lo w	
2	l o w e r	l o w e r	l o w e r	lo w e r	
1	l o w e s t	l o w es t	l o w est	lo w est	・・・
6	n e w e s t	n e w es t	n e w est	n e w est	
3	w i d e s t	w i d es t	w i d est	w i d est	

Model Architecture

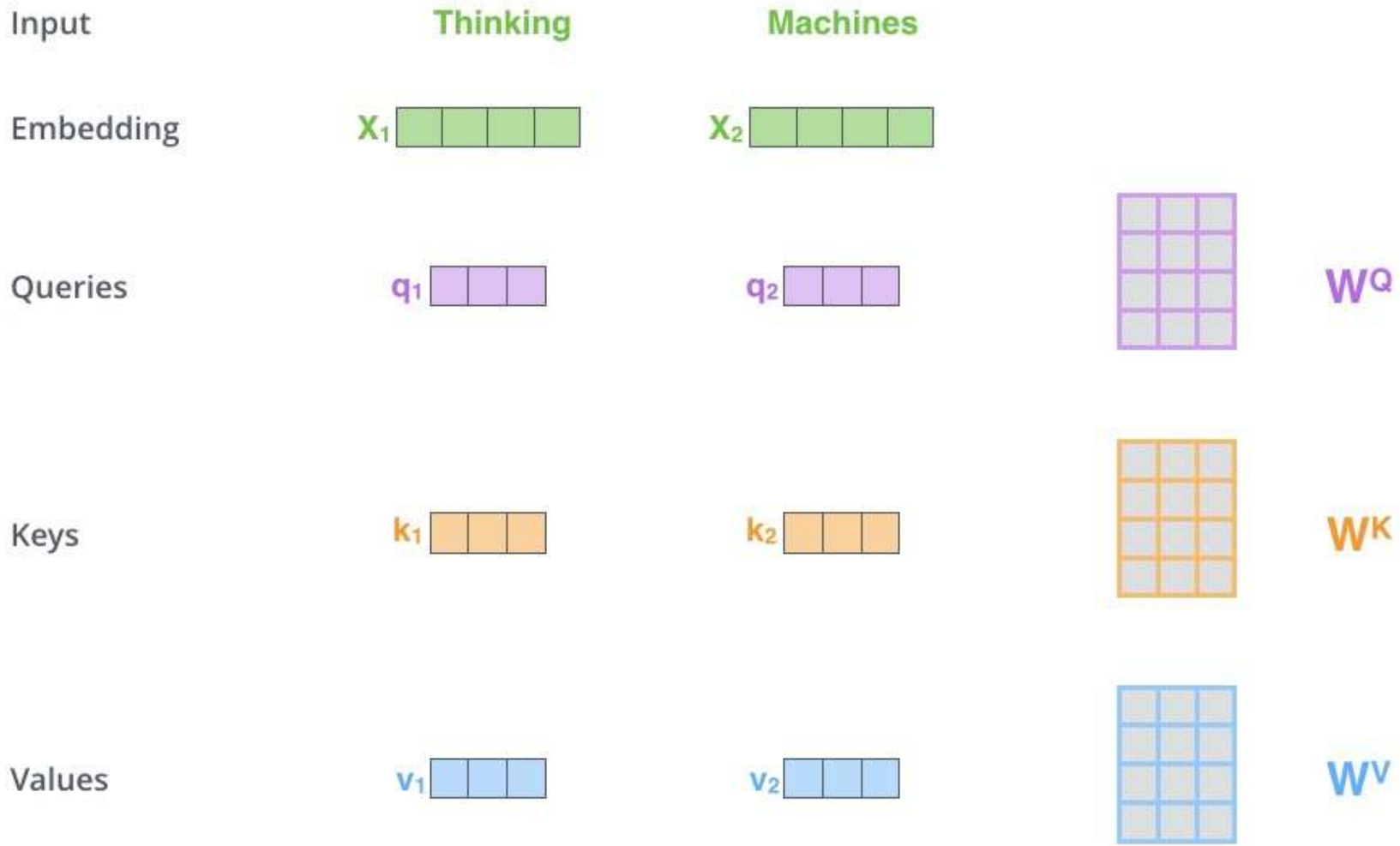




Transformer [Vaswani+ 2017]

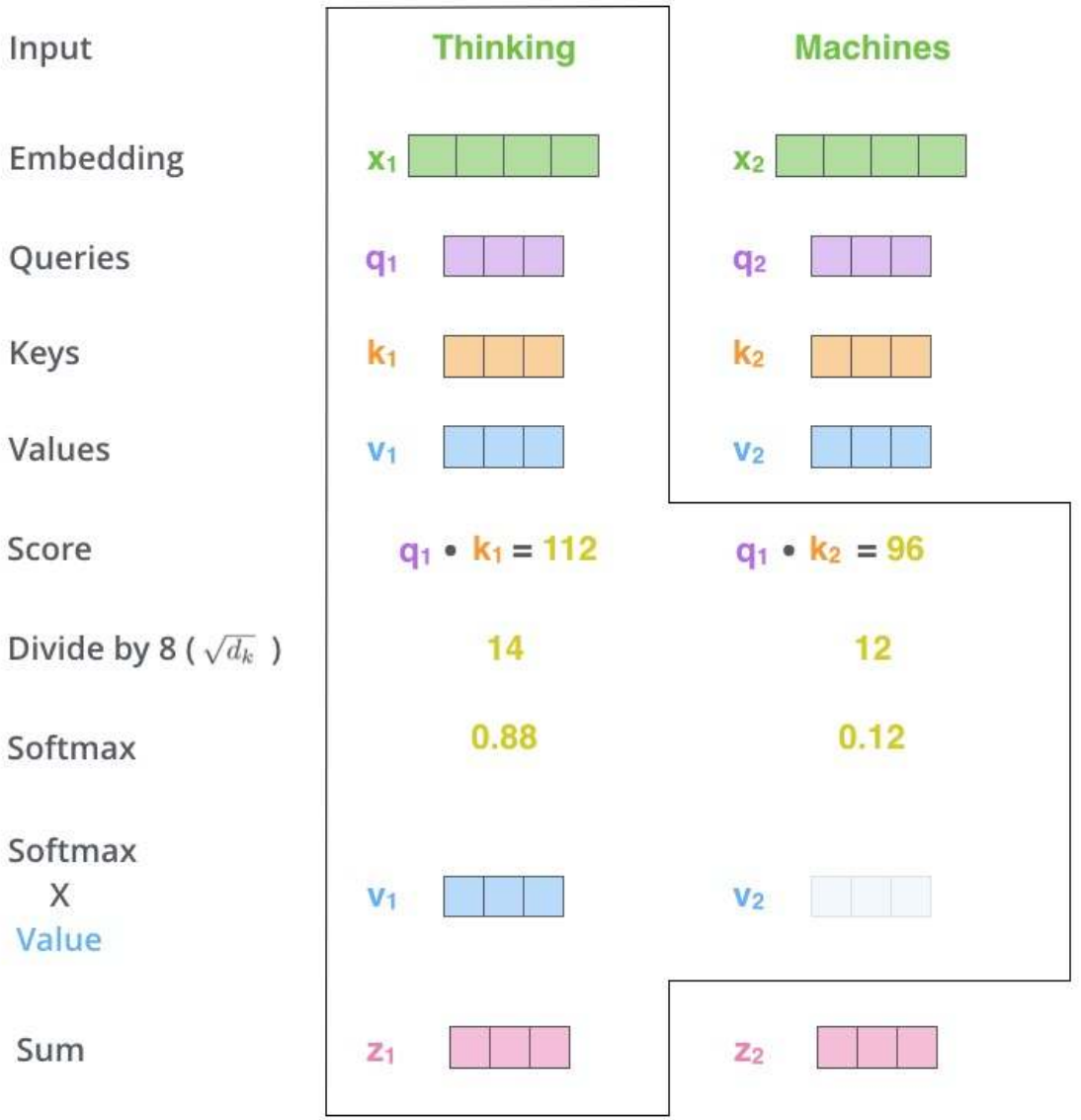
The Annotated Transformer:
http://nlp.seas.harvard.edu/2018/04/03/attention.html

Query, Key, Value

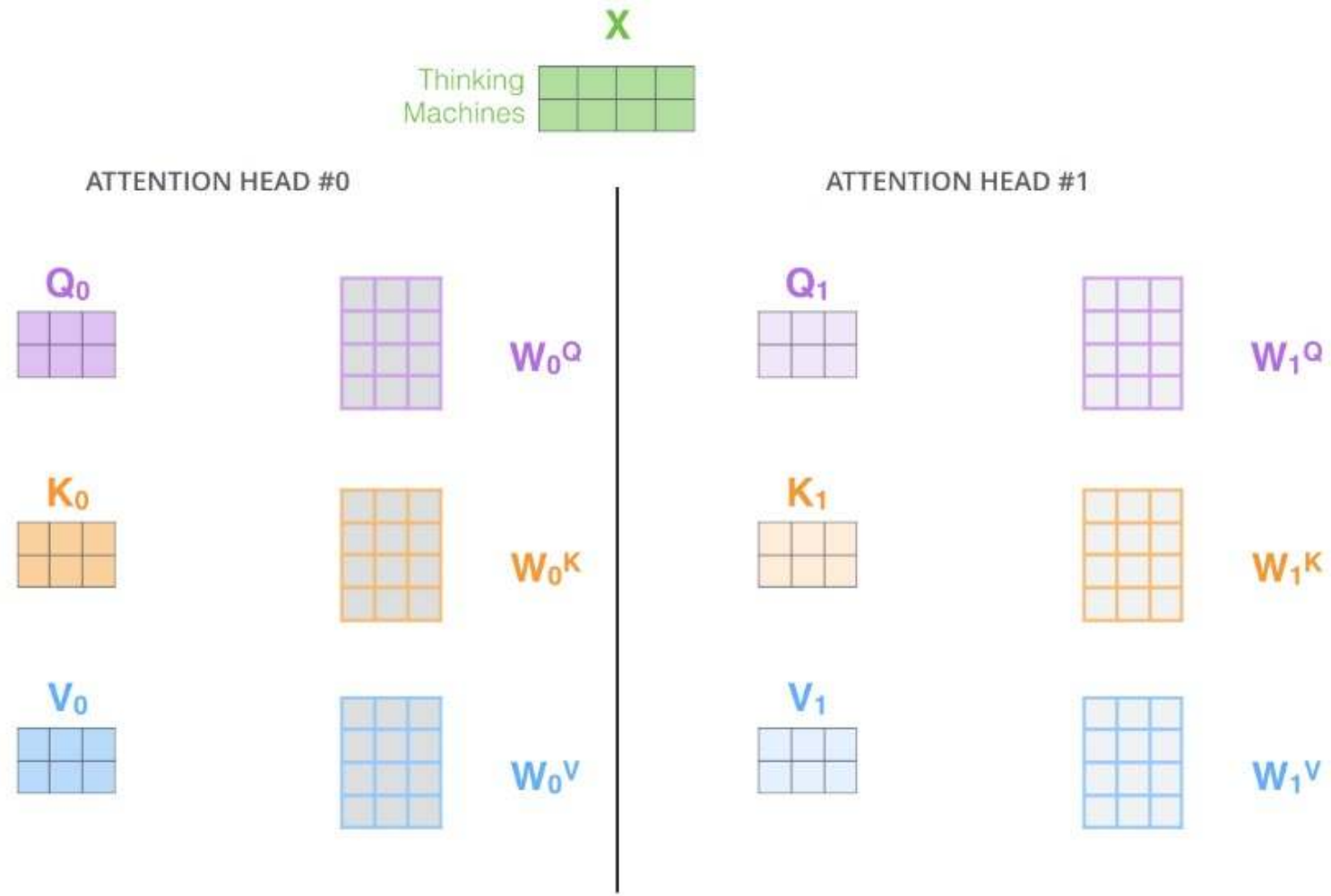


https://jalammar.github.io/illustrated-transformer/

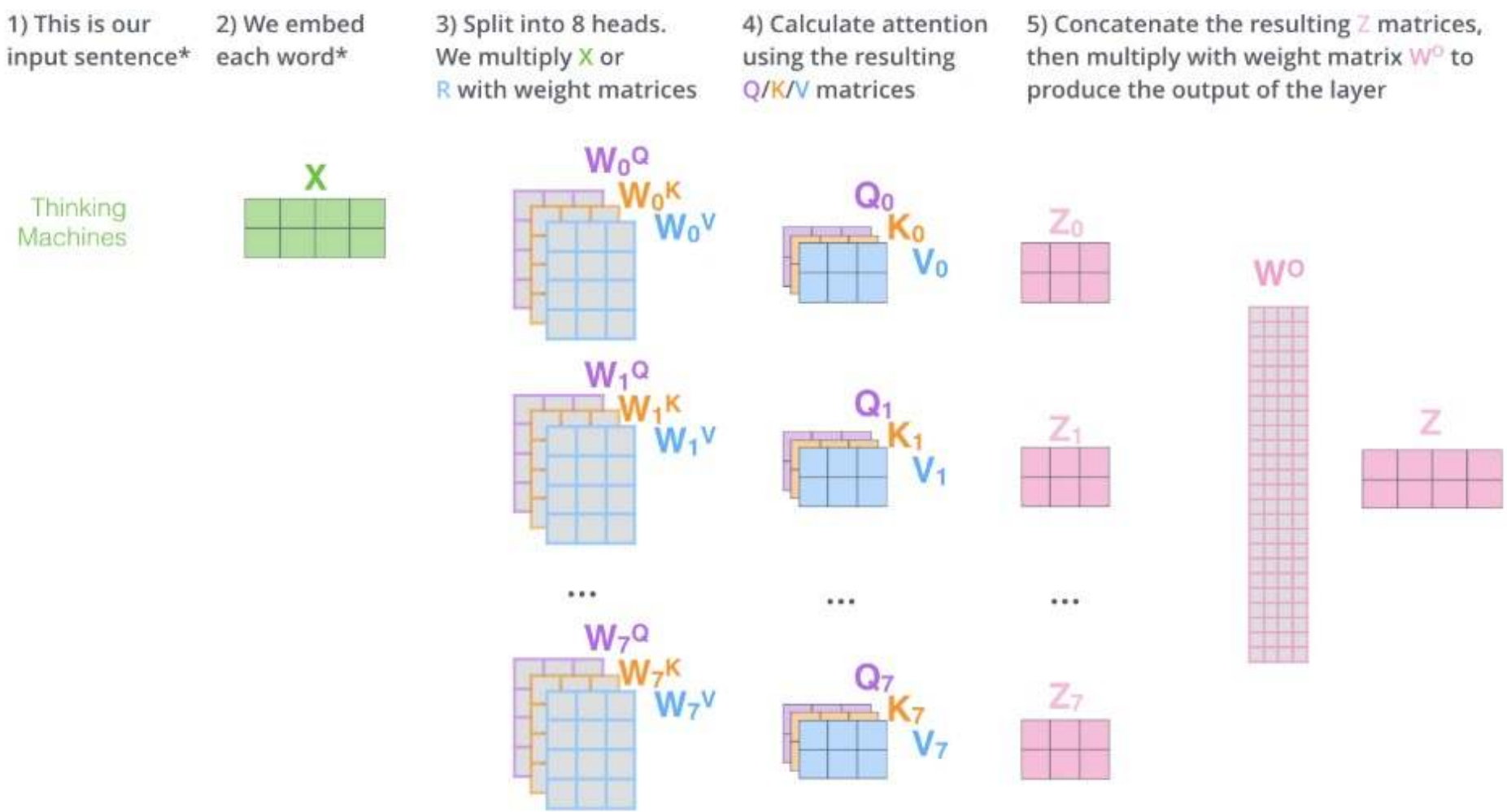
Self-Attention



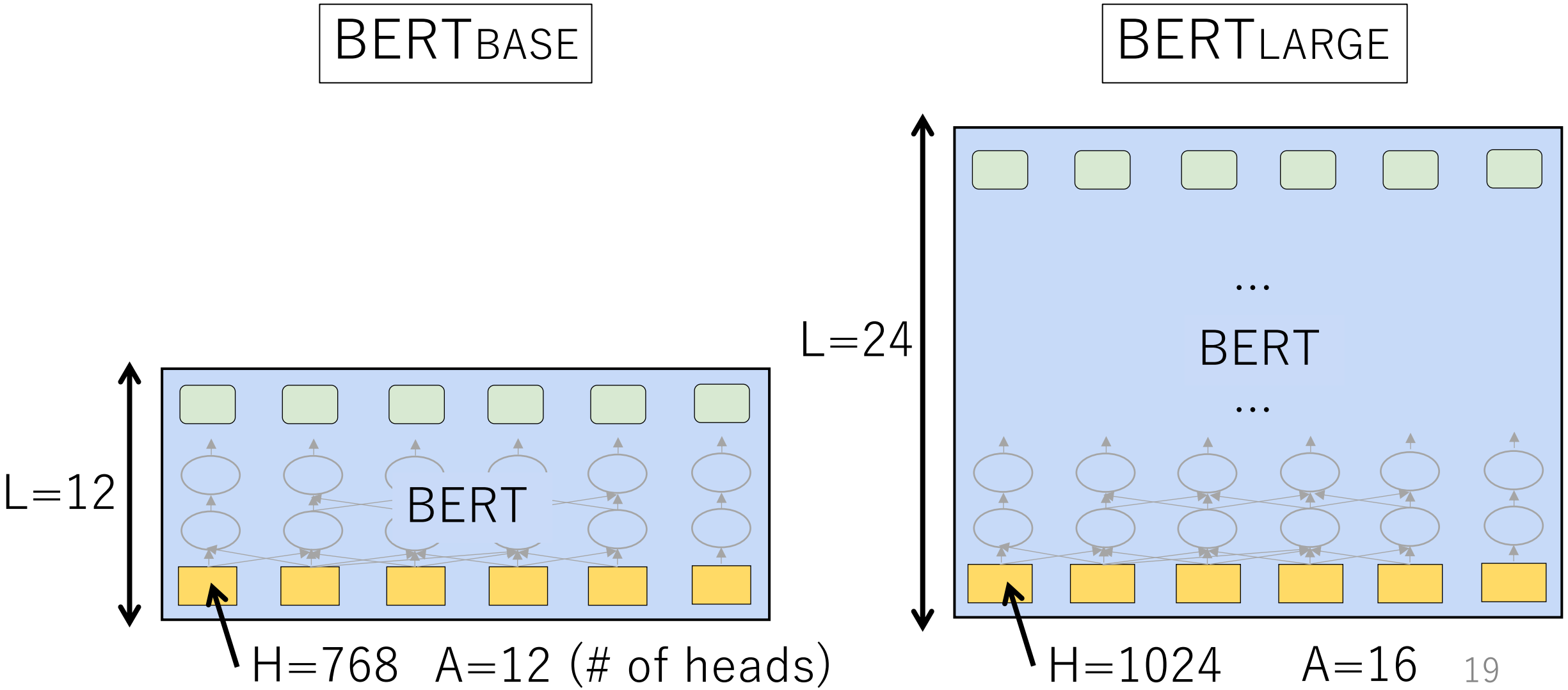
Multiple Heads



Summary

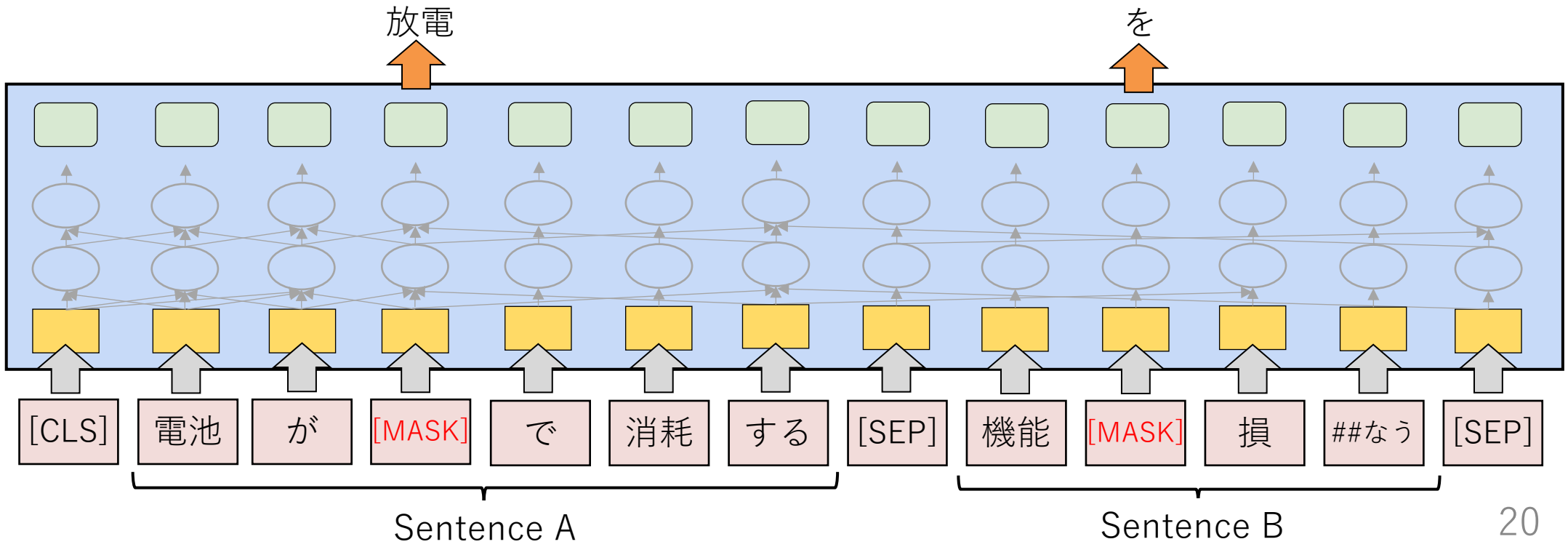

1) This is our input sentence*
2) We embed each word*
3) Split into 8 heads. We multiply X or R with weight matrices
4) Calculate attention using the resulting Q/K/V matrices
5) Concatenate the resulting Z matrices, then multiply with weight matrix W^O to produce the output of the layer
Thinking
Machines
X
W0Q
W0K
W0V
Q0
K0
V0
Z0
W1Q
W1K
W1V
Q1
K1
V1
Z1
...
...
...
W7Q
W7K
W7V
Q7
K7
V7
Z7
WO
Z

Models


BERTBASE
BERTLARGE
L=12
BERT
H=768 A=12 (# of heads)
L=24
...
BERT
...
H=1024 A=16

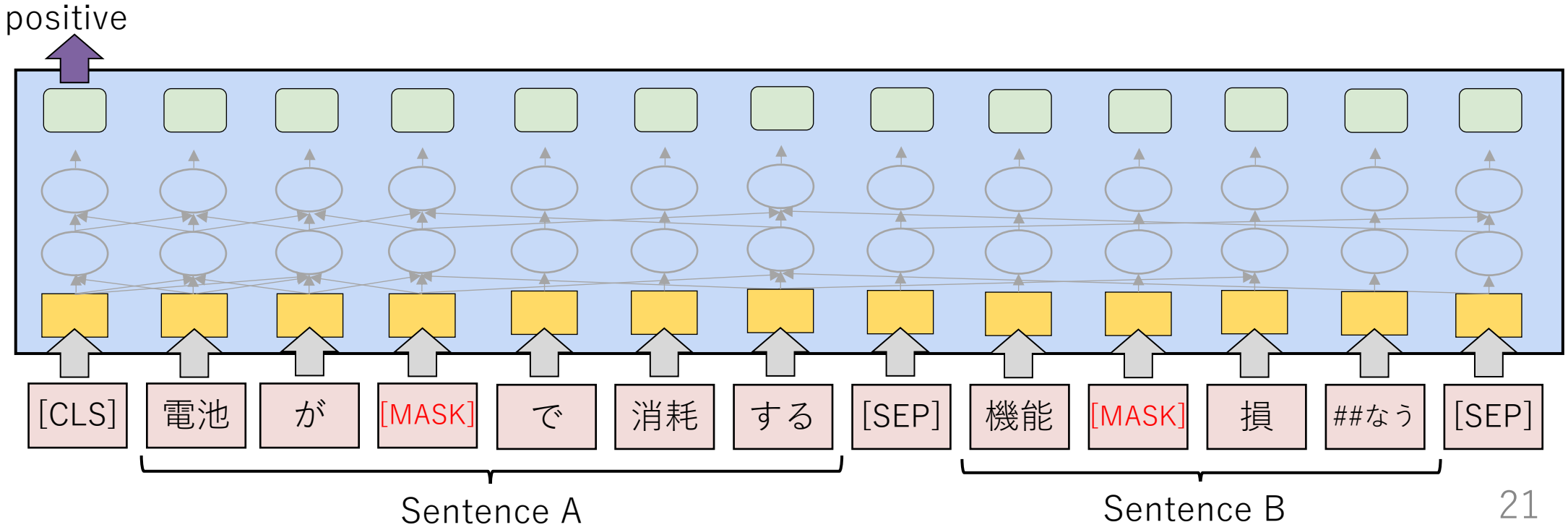
Pre-training: 1. Masked Language Model

- Mask a randomly selected token in an input sentence
- Predict the masked token from its context
  - It is required to learn a syntactic and semantic representation



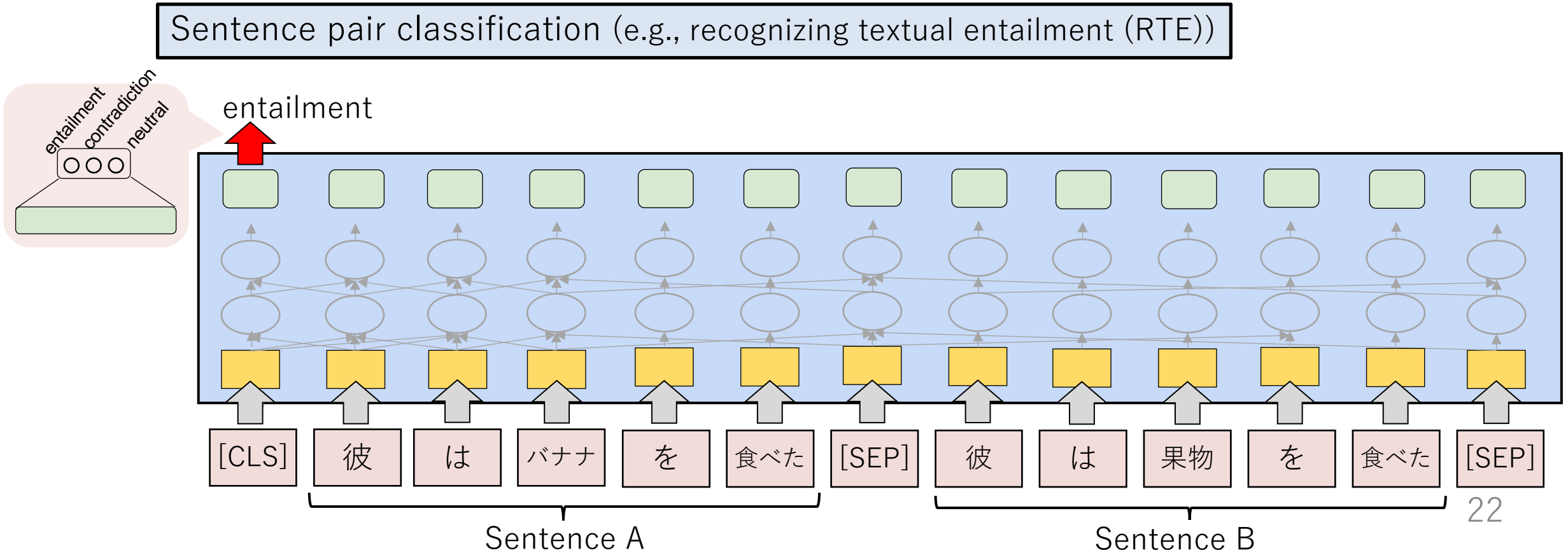
Pre-training: 2. Next Sentence Prediction

- It is necessary to capture relations between two sentences for entailment recognition and question answering
- Solve a problem to distinguish sentence B following sentence A from a randomly selected sentence



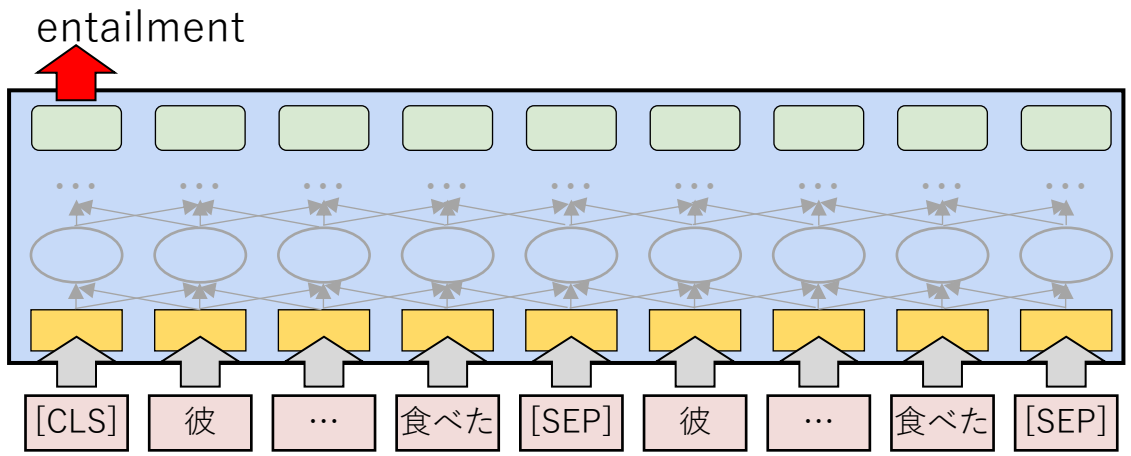
Fine-Tuning (1/2)

- Add a top layer according to a downstream task
- Update the parameters of the top layer and the Transformer

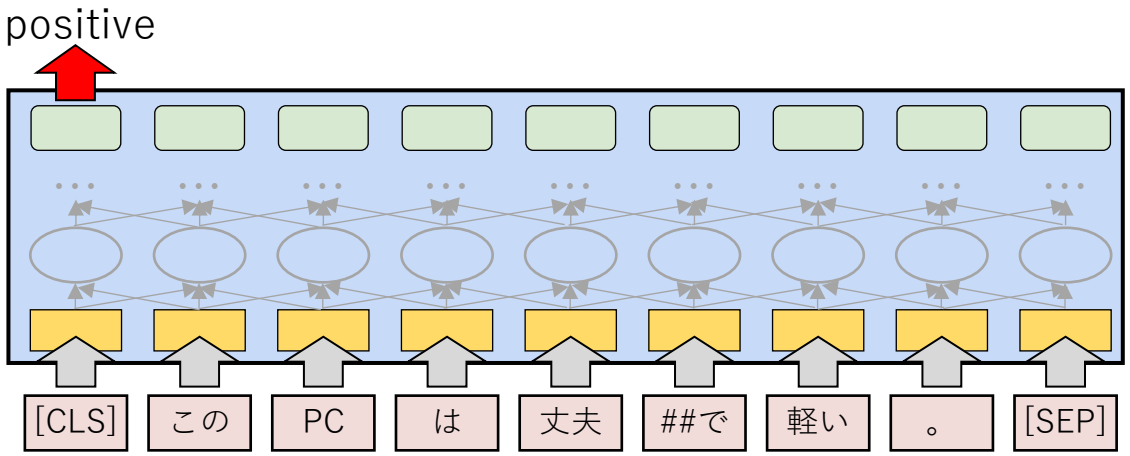


Fine-Tuning (2/2)

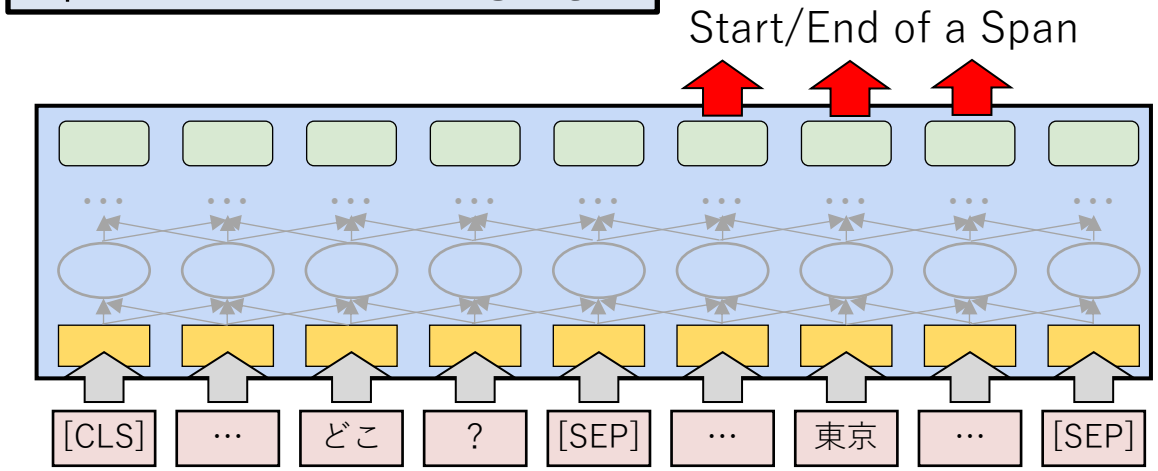
Sentence pair classification (e.g., RTE)



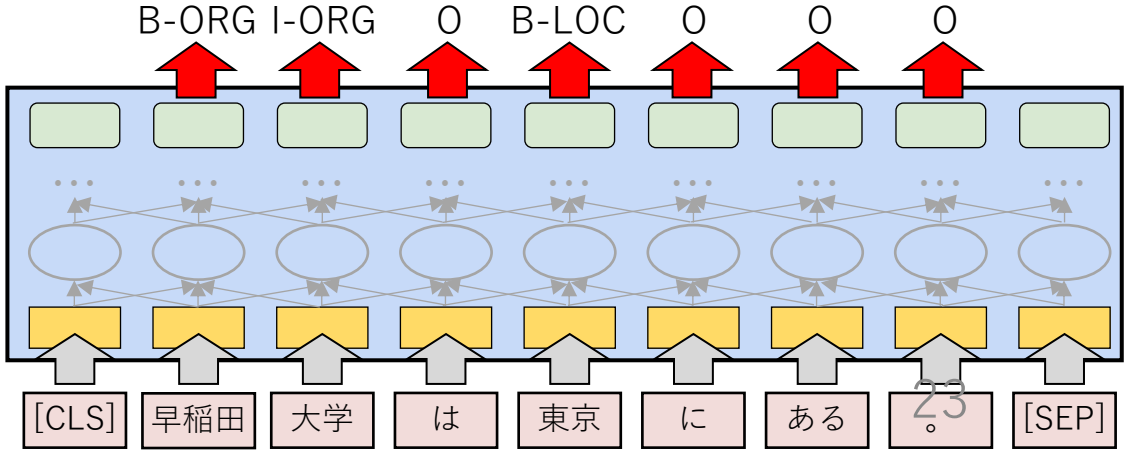
Sentence classification (e.g., sentiment analysis)



Span extraction (e.g., QA)



Sequence labeling (e.g., NER)



GLUE (General Language Understanding Evaluation)

[Wang+ 18]

	Task	Description
1s	SST-2	Polarity (positive/negative) of movie reviews
1s	CoLA	Whether a sentence is acceptable
2s	MRPC	Whether 2 sentences have the same meaning
2s	STS-B	Similarity of 2 sentences (1-5)
2s	QQP	Whether 2 question sentences have the same meaning
2s	MNLI	2 sentences have a relation of entailment/contradiction/neutral
2s	QNLI	Whether a sentence contains an answer for a question (from SQuAD)
2s	RTE	Whether 2 sentences have an entailment relation
2s	WNLI	Whether 2 sentences have an entailment relation (from Winograd Schema Challenge)

GLUE (General Language Understanding Evaluation)

[Wang+ 18]

[positive]
it 's a charming and often affecting journey .

[negative]
a sometimes tedious film .

Task	Description
SST-2	Polarity (positive/negative) of movie reviews
CoLA	Whether a sentence is acceptable
MRPC	Whether 2 sentences have the same meaning
STS-B	Similarity of 2 sentences (1-5)

1s

2s

[same]
The top rate will go to 4.45 percent for all residents with taxable incomes above $ 500,000 .
For residents with incomes above $ 500,000 , the income-tax rate will increase to 4.45 percent .

[not same]
While dioxin levels in the environment were up last year , they have dropped by 75 percent since the 1970s , said Caswell .
The Institute said dioxin levels in the environment have fallen by as much as 76 percent since the 1970s .

https://gluebenchmark.com/leaderboard

GLUE Leaderboard (as of November 2019)

T5: 89.7

Human performance: 87.1

BERT: 80.5

Baseline (ELMo): 70.0

	Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
	1	T5 Team - Google	T5		89.7	70.8	97.1	91.9/89.2	92.5/92.1	74.6/90.4	92.0	91.7	96.7	92.5	93.2	53.1
	2	ALBERT-Team Google Language	ALBERT (Ensemble)		89.4	69.1	97.1	93.4/91.2	92.5/92.0	74.2/90.5	91.3	91.0	99.2	89.2	91.8	50.2
+	3	王玮	ALICE v2 large ensemble (Alibaba DAMO NLP)		89.0	69.2	97.1	93.6/91.5	92.7/92.3	74.4/90.7	90.7	90.2	99.2	87.3	89.7	47.8
	4	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.8	68.0	96.8	93.1/90.8	92.4/92.2	74.8/90.3	91.1	90.7	98.8	88.7	89.0	50.1
	5	Facebook AI	RoBERTa		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	88.2	89.0	48.7
	6	XLNet Team	XLNet-Large (ensemble)		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	86.3	90.4	47.5
+	7	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	[illegible]	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
	8	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	-
	9	Stanford Hazy Research	Snorkel MeTaL		83.2	63.8	96.2	91.5/88.5	90.1/89.7	73.1/89.9	87.6	87.2	93.9	80.9	65.1	39.9
	10	XLM Systems	XLM (English only)		83.1	62.9	95.6	90.7/87.1	88.8/88.2	73.2/89.8	89.1	88.5	94.0	76.0	71.9	44.7
	11	Zhuosheng Zhang	SemBERT		82.9	62.3	94.6	91.2/88.3	87.8/86.7	72.8/89.8	87.6	86.3	94.6	84.5	65.1	42.4
	12	Danqi Chen	SpanBERT (single-task training)		82.8	64.3	94.8	90.9/87.9	89.9/89.1	71.9/89.5	88.1	87.7	94.3	79.0	65.1	45.1
	13	Kevin Clark	BERT + BAM		82.3	61.5	95.2	91.3/88.3	88.6/87.9	72.5/89.7	86.6	85.8	93.1	80.4	65.1	40.7
	14	Nitish Shirish Keskar	Span-Extractive BERT on STILTs		[illegible]	[illegible]	[illegible]	[illegible]	89.4/89.2	72.2/89.4	86.5	85.8	92.5	79.8	65.1	28.3
	15	Jason Phang	BERT on STILTs		[illegible]	[illegible]	[illegible]	[illegible]	88.7/88.3	71.9/89.4	86.4	85.6	92.7	80.1	65.1	28.3
	16	廖亿	RGLM-Base (Huawei Noah's Ark Lab)		[illegible]	56.9	94.2	90.7/87.7	89.7/89.1	72.2/89.4	86.1	85.4	92.1	78.5	65.1	40.0
+	17	Jacob Devlin	BERT: 24-layers, 16-heads, 1024-hidden		80.5	60.5	94.9	89.3/85.4	87.6/86.5	72.1/89.3	86.7	85.9	92.7	70.1	65.1	39.6
	18	Neil Houlsby	BERT + Single-task Adapters		80.2	59.2	94.3	88.7/84.3	87.3/86.1	71.5/89.4	85.4	85.0	92.4	71.6	65.1	9.2
	19	Zhuohan Li	Macaron Net-base		79.7	57.6	94.0	88.4/84.4	87.5/86.3	70.8/89.0	85.4	84.5	91.6	70.5	65.1	38.7
	20	蘇大鈞	SesameBERT-Base		78.6	52.7	94.2	88.9/84.8	86.5/85.5	70.8/88.8	83.7	83.6	91.0	67.6	65.1	35.8
+	21	MobileBERT Team	MobileBERT		78.5	51.1	92.6	88.8/84.5	86.2/84.8	70.5/88.3	84.3	83.4	91.6	70.4	65.1	34.3
	22	Linyuan Gong	StackingBERT-Base		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	84.4	84.2	90.1	67.0	65.1	36.6
	23	Huawei Noah's Ark Lab	TinyBERT (4-layers; 7.5x smaller and 9.4x faster than BERT-ba		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	82.5	81.8	87.7	62.9	65.1	33.7
	24	shijing si	bert+pos6		[illegible]	52.9	93.9	88.8/84.6	83.8/85.5	71.4/89.2	84.4	83.3	90.4	66.9	34.9	0.0
	25	GLUE Baselines	BiLSTM+ELMo+Attn		70.0	33.6	90.4	84.4/78.0	74.2/72.3	63.1/84.3	74.1	74.5	79.8	58.9	65.1	21.7

Pre-training for Japanese

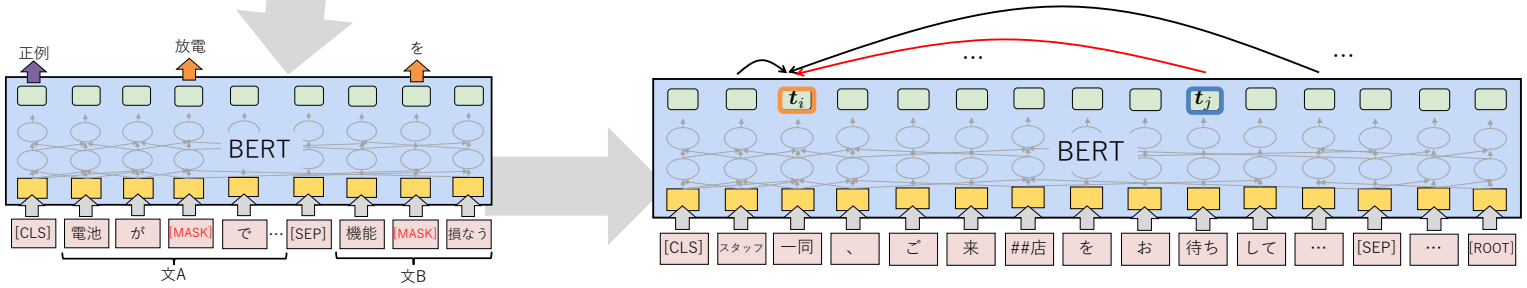
Japanese Wikipedia
(18M sentences)

自動車 競技 は 四 輪 の 自動車 あるいは それ に 準ずる 車両 による 競技 に 対して 主に 呼称 され、オートバイ や それ に 準ずる 車両 の 競技 に 対して は オートバイ 競技 や モーター サイクル レース など と 呼ば れる 。自動車 競技 は操る 人 の …
…
ＰＣＭ 音源 （ ピー ##シー ##エム おん ##げん ） は 、 コンパクト ディスク など で 扱わ れる パルス 符号 変調 技術 を 用いた デジタル ##シンセサイザー の 音源 方式 の ひと つ 。 あらかじめ メモリ に 記録 して おいた ＰＣＭ 波形 （ サンプル ） を 再生 する こと で …
…

Data for pre-training

[CLS] 自動車 競技 は 四 輪 の 自動車 あるいは それ に 準ずる 車両 に よる 競技 に 対して 主に 呼称 さ れ 、 オートバイ [MASK] … 呼ばれる 。 [SEP] 自動車 競技 は操る [MASK] の… [SEP]
Label: IsNext

Pre-training

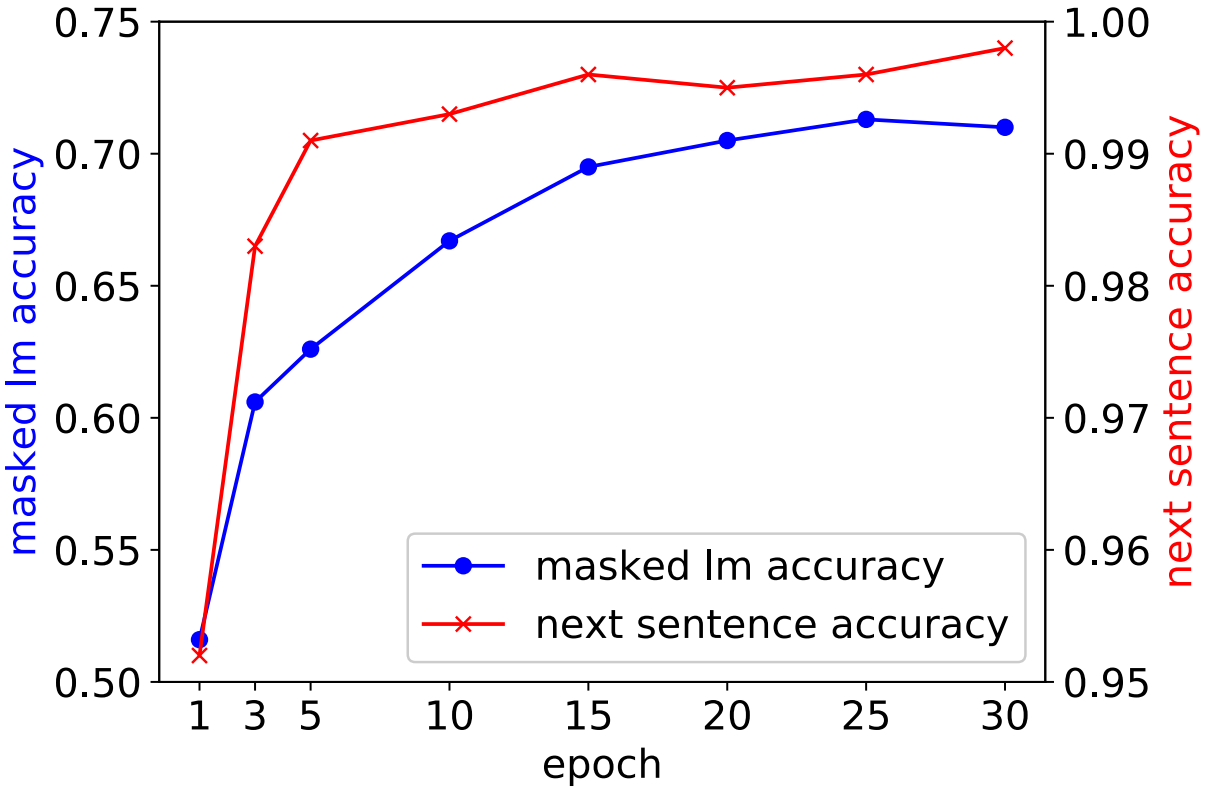


Fine-tuning

Using morphological analysis and BPE

Pre-training for Japanese

- Vocab size: 32000, BERT$_{\text{BASE}}$, max token number: 128



1 day for 1 epoch
(using 1 GPU)

BERT Masked LM

京都市左京区の京都大学では１７日、試験について案内する看板の設置作業が行われた。

input	predictions
京都	京都, 同, 大阪, 福岡, 東京
市	市, 府, ・, 都, 大学
左京	左京, 東山, 中京, 右京, 伏見
区	区, 部, 地区, ##区, 地域
の	の, と, および, 、, , ・
京都	京都, 立命館, 同志社, 近畿, 龍谷
大学	大学, 駅, 支所, キャンパス, 支社
で	で, に, へ, にて, と
は	は, 、, , 翌, 同, 毎月
１７	この, その, ある, １, ２４
日	日, 日間, 月, 年間, 年
、	に, から, 、, , より, まで
試験	大学, キャンパス, 学生, これ, 授業
に	に, ##に, 結果, 詳細に, 的に
ついて	ついて, 向けて, おいて, 際して, 合わせて
案内	説明, 告知, 案内, 解説, 掲示
する	する, 用, 表示, の, 説明
看板	板, 看板, 装置, 掲示板, ため
の	の, を, と, や, が
設置	設置, 撤去, 取り付け, 作成, 貼り
作業	作業, 式, 工事, など, 実験
が	が, も, を, まで, で
行わ	行わ, 行なわ, おこなわ, さ, 開か
れた	れた, れる, れ, られた, れて
。	。, 、, , に, も, ため

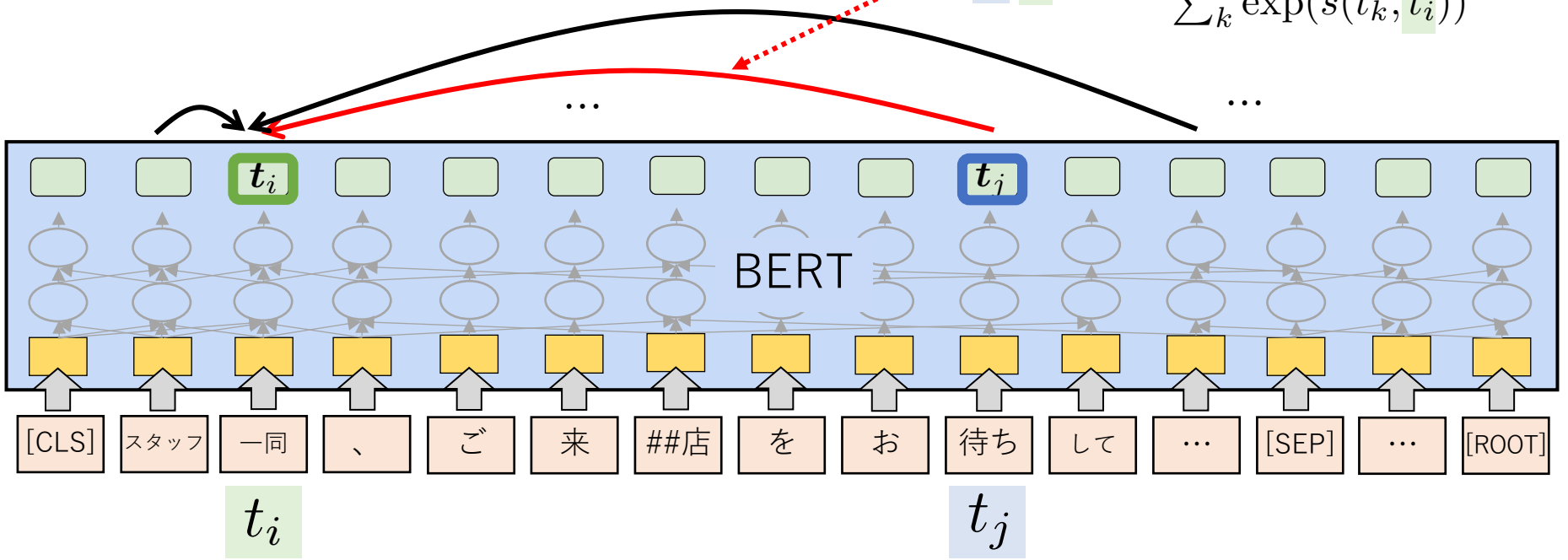
BERT Fine-tuning for Dependency Parsing

[柴田+ 19]

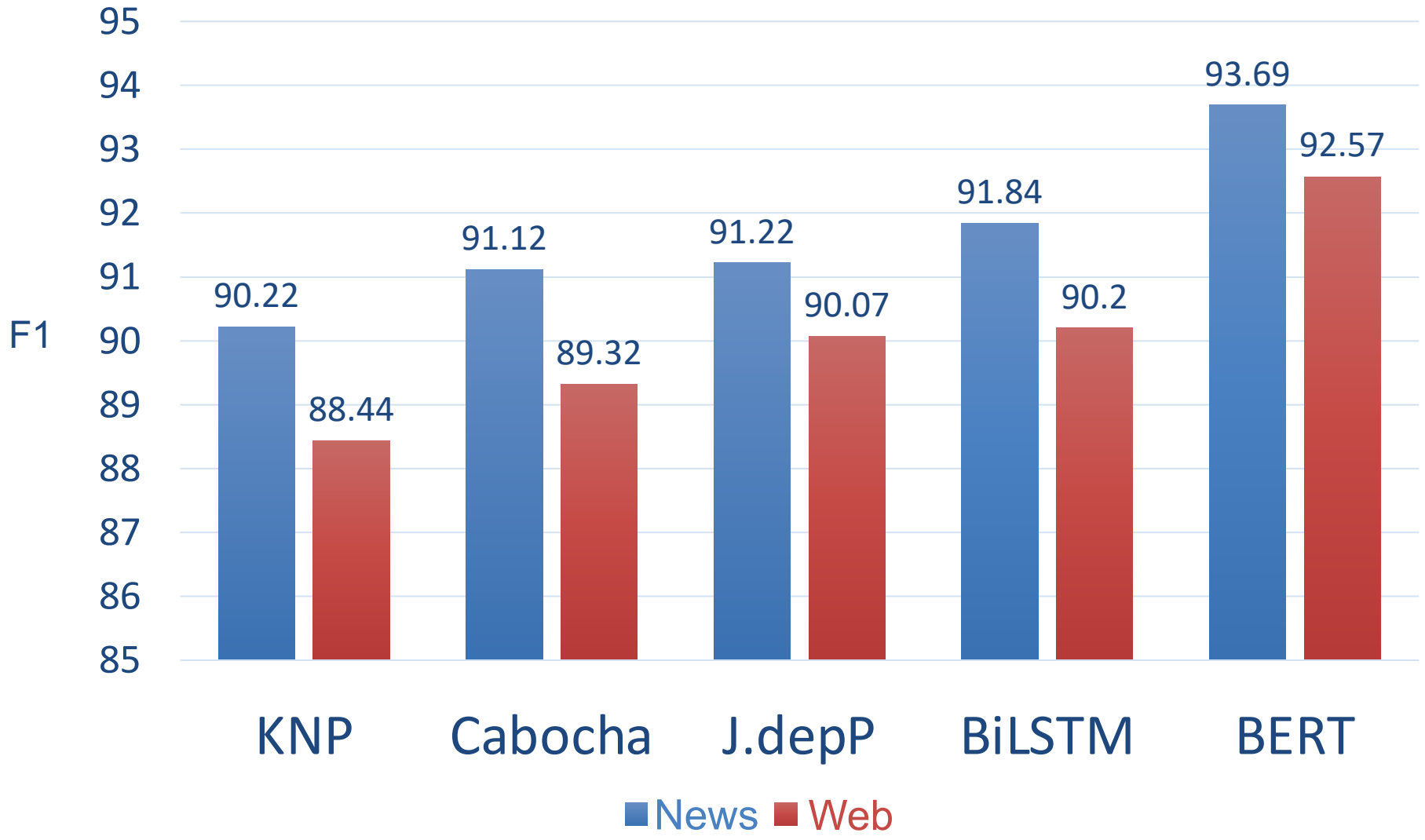
Based on
head selection [Zhang+ 17]

$$s(t_j, t_i) = \boldsymbol{v}_h^{\mathrm{T}} \tanh(U_h \boldsymbol{t}_j + W_h \boldsymbol{t}_i)$$

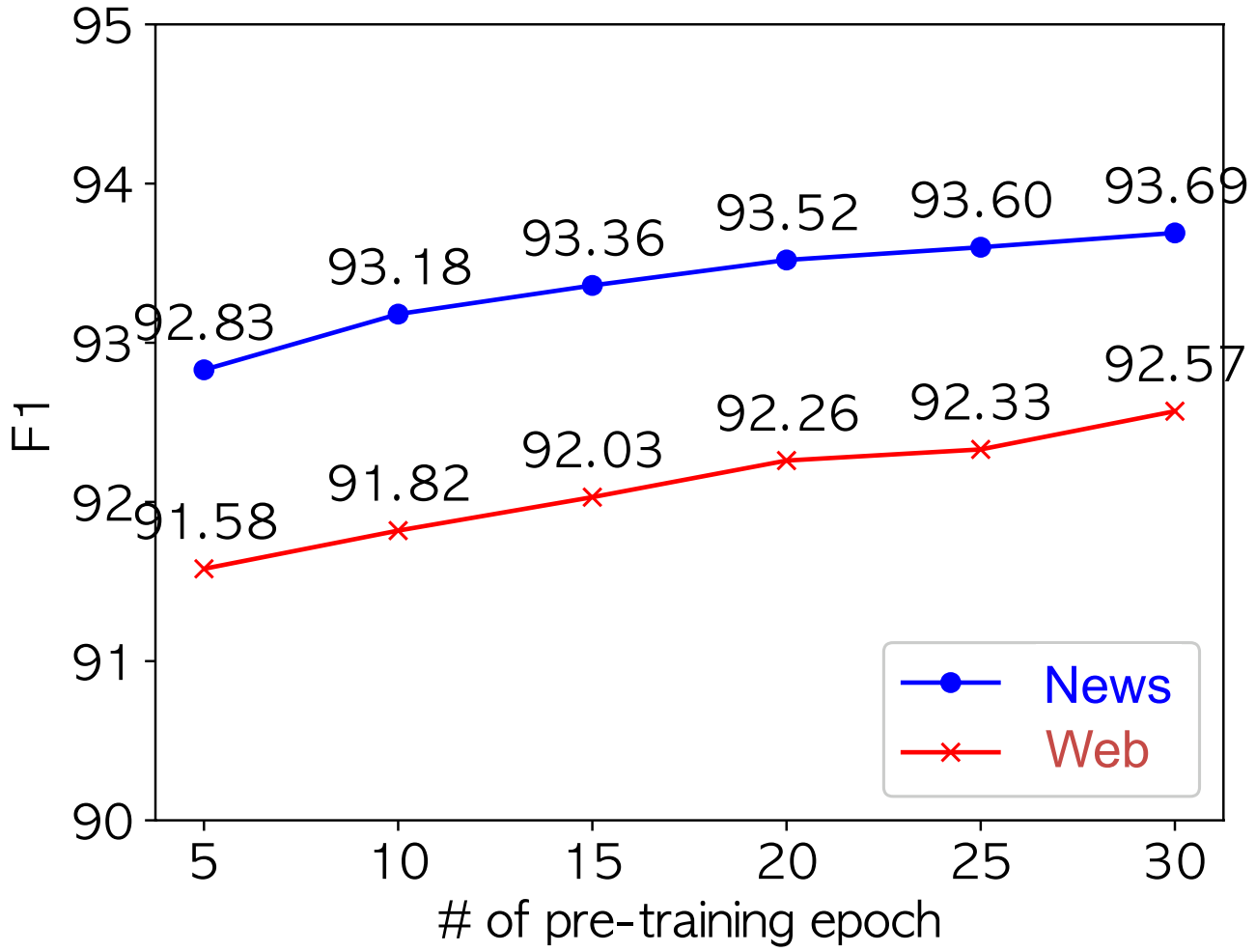
$$P_{head}(t_j | t_i, S) = \frac{\exp(s(t_j, t_i))}{\sum_k \exp(s(t_k, t_i))}$$



Phrase-based Performance (F1)



Learning Curve while Increasing Epochs of Pre-training



Recent Topics around BERT

- Model improvements
- Use of knowledge
- Language models
- Generation models for translation, summarization, etc.
- Multilingualization
- Analysis of inside BERT
- Use of multimodal information

RoBERTa [Liu+ 2019] (FAIR&UW)

A Robustly Optimized BERT Pretraining Approach

- Dynamic masking
- One-segment pre-training without NSP
- Larger batch size (32x BERT = 8K)
- Larger corpora (10x BERT = 160GB)
- Longer pre-training steps (16x BERT)
- Byte-level BPE
  - larger vocabulary (50K subwords)

RoBERTa [Liu+ 2019]

Model	data	bsz	steps	SQuAD (v1.1/2.0)	MNLI-m	SST-2
RoBERTa						
with BOOKS + WIKI	16GB	8K	100K	93.6/87.3	89.0	95.3
+ additional data (§3.2)	160GB	8K	100K	94.0/87.7	89.3	95.6
+ pretrain longer	160GB	8K	300K	94.4/88.7	90.0	96.1
+ pretrain even longer	160GB	8K	500K	**94.6/89.4**	**90.2**	**96.4**
$\text{BERT}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	90.9/81.8	86.6	93.7
$\text{XLNet}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	94.0/87.8	88.4	94.4
+ additional data	126GB	2K	500K	94.5/88.8	89.8	95.6

RoBERTa [Liu+ 2019]

GLUE

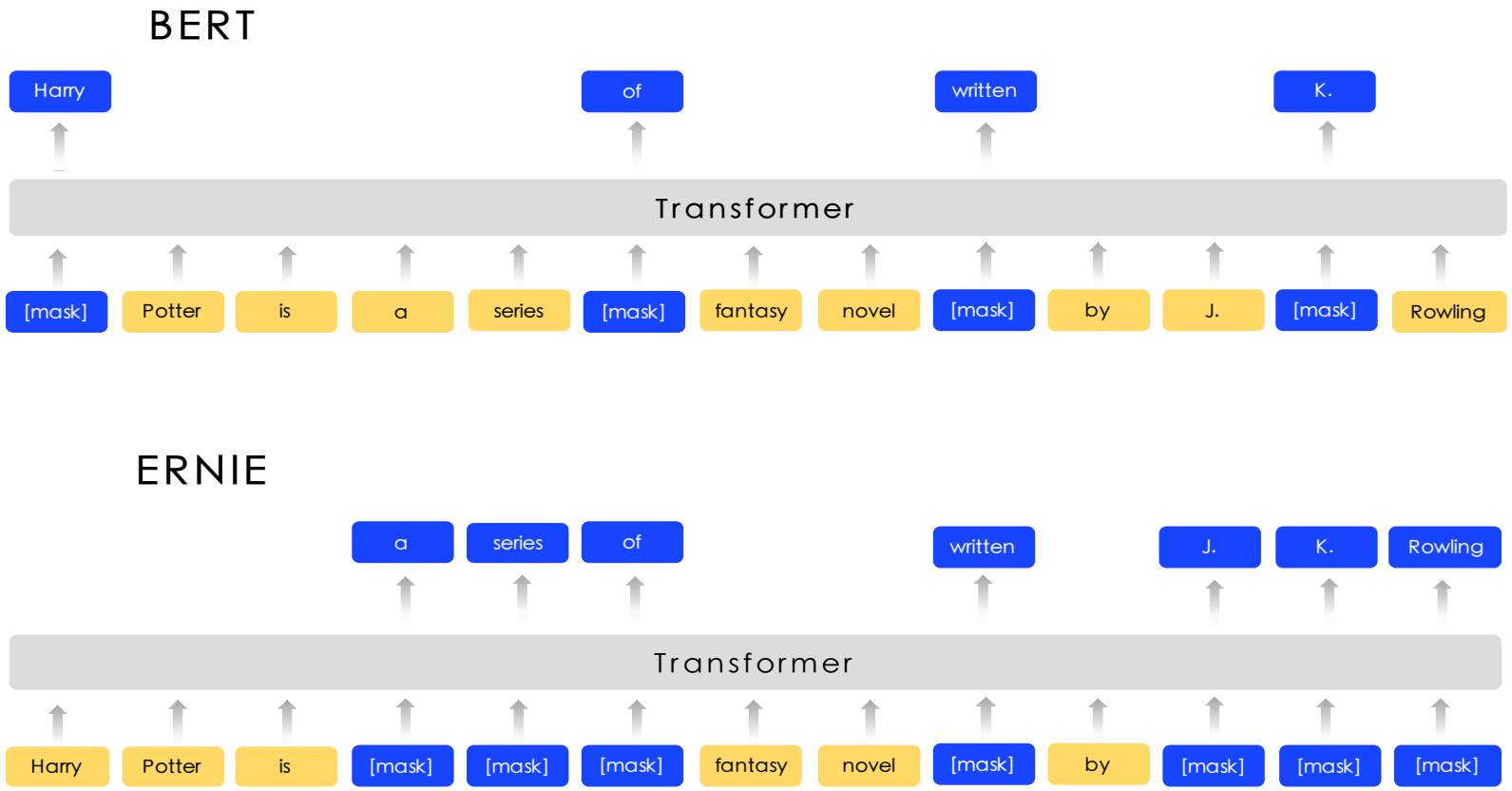
	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS	WNLI	Avg
Single-task single models on dev										
$\text{BERT}_{\text{LARGE}}$	86.6/-	92.3	91.3	70.4	93.2	88.0	60.6	90.0	-	-
$\text{XLNet}_{\text{LARGE}}$	89.8/-	93.9	91.8	83.8	95.6	89.2	63.6	91.8	-	-
RoBERTa	**90.2/90.2**	**94.7**	**92.2**	**86.6**	**96.4**	**90.9**	**68.0**	**92.4**	**91.3**	-
Ensembles on test (from leaderboard as of July 25, 2019)										
ALICE	88.2/87.9	95.7	**90.7**	83.5	95.2	92.6	**68.6**	91.1	80.8	86.3
MT-DNN	87.9/87.4	96.0	89.9	86.3	96.5	92.7	68.4	91.1	89.0	87.6
XLNet	90.2/89.8	98.6	90.3	86.3	**96.8**	**93.0**	67.8	91.6	**90.4**	88.4
RoBERTa	**90.8/90.2**	**98.9**	90.2	**88.2**	96.7	92.3	67.8	**92.2**	89.0	**88.5**

Whole Word Masking (WWM)

- Input Text:
  the man jumped up , put his basket on phil ##am ##mon ' s head
- Original Masked Input:
  [MASK] man [MASK] up , put his [MASK] on **phil [MASK] ##mon** ' s head
- Whole Word Masked Input:
  the man [MASK] up , put his basket on **[MASK] [MASK] [MASK]** ' s head

ERNIE [Sun+ 2019] (Baidu)

- Phrase-level masking (using chunking)
- Entity-level masking (using NER)



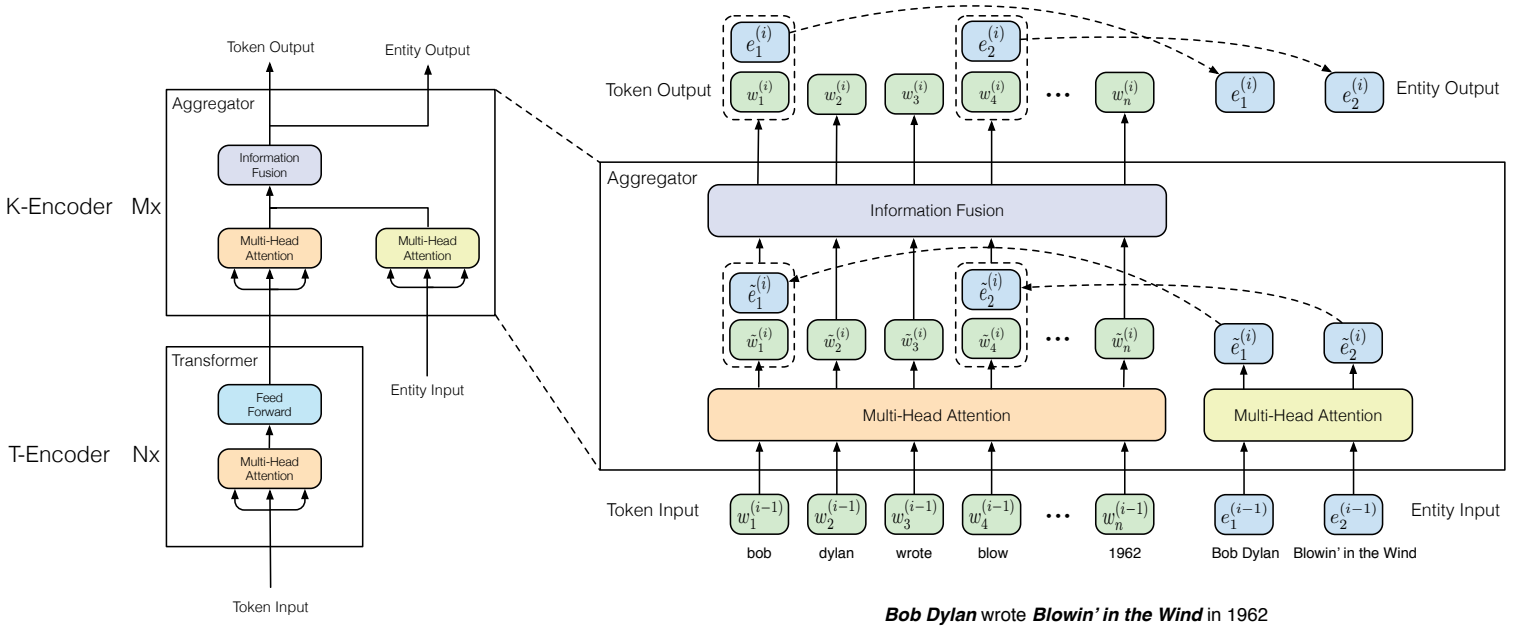
ERNIE [Sun+ 2019]

Table 1: Results on 5 major Chinese NLP tasks

Task	Metrics	Bert		ERNIE	
		dev	test	dev	test
XNLI	accuracy	78.1	77.2	79.9 (+1.8)	78.4 (+1.2)
LCQMC	accuracy	88.8	87.0	89.7 (+0.9)	87.4 (+0.4)
MSRA-NER	F1	94.0	92.6	95.0 (+1.0)	93.8 (+1.2)
ChnSentiCorp	accuracy	94.6	94.3	95.2 (+0.6)	95.4 (+1.1)
nlpcc-dbqa	mrr	94.7	94.6	95.0 (+0.3)	95.1 (+0.5)
	F1	80.7	80.8	82.3 (+1.6)	82.7 (+1.9)

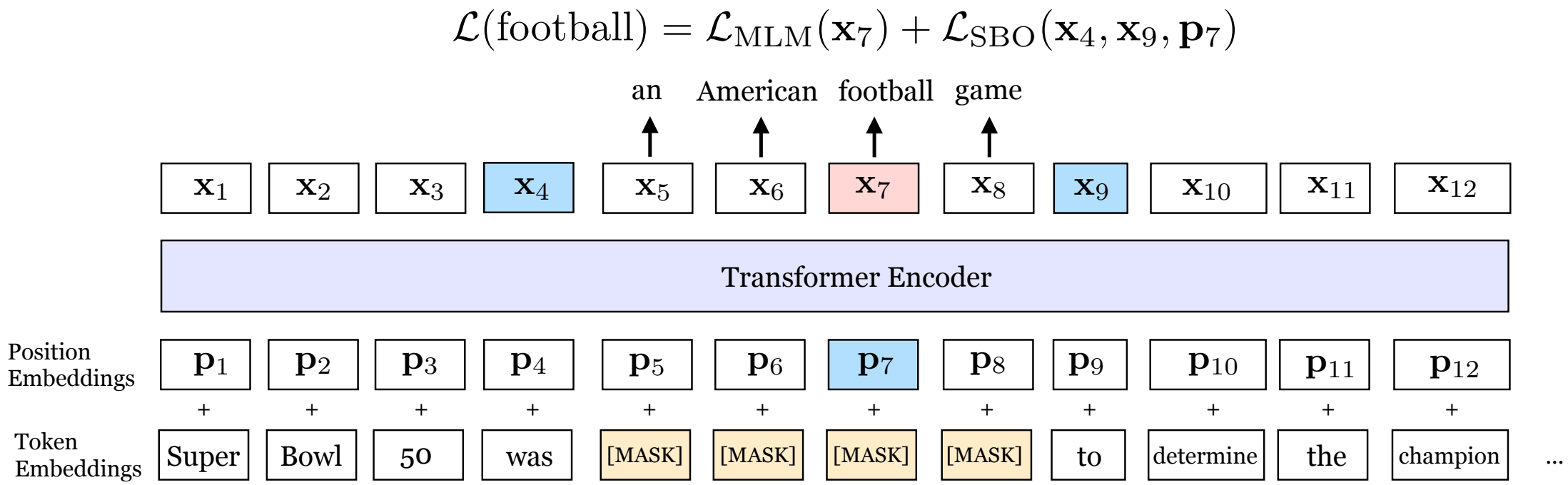
ERNIE [Zhang+ 2019] (Tsinghua)

- Pre-training to predict entities
- Knowledgeable encoder
  - Entity embeddings (induced by TransE) are integrated



SpanBERT [Joshi+ 2019] (FAIR&UW)

- Span-level masking
- Span Boundary Objective (SBO)
- One-segment pre-training without NSP



SpanBERT [Joshi+ 2019]

	SQuAD 1.1		SQuAD 2.0	
	EM	F1	EM	F1
Human Perf.	82.3	91.2	86.8	89.4
Google BERT	84.3	91.3	80.0	83.3
Our BERT	86.5	92.6	82.8	85.9
Our BERT-1seq	87.5	93.3	83.8	86.6
SpanBERT	**88.8**	**94.6**	**85.7**	**88.7**

	NewsQA	TriviaQA	SearchQA	HotpotQA	NaturalQA	(Avg)
Google BERT	68.8	77.5	81.7	78.3	79.9	77.3
Our BERT	71.0	79.0	81.8	80.5	80.5	78.6
Our BERT-1seq	71.9	80.4	84.0	80.3	81.8	79.7
SpanBERT	**73.6**	**83.6**	**84.8**	**83.0**	**82.5**	**81.5**

SpanBERT [Joshi+ 2019]

Coreference

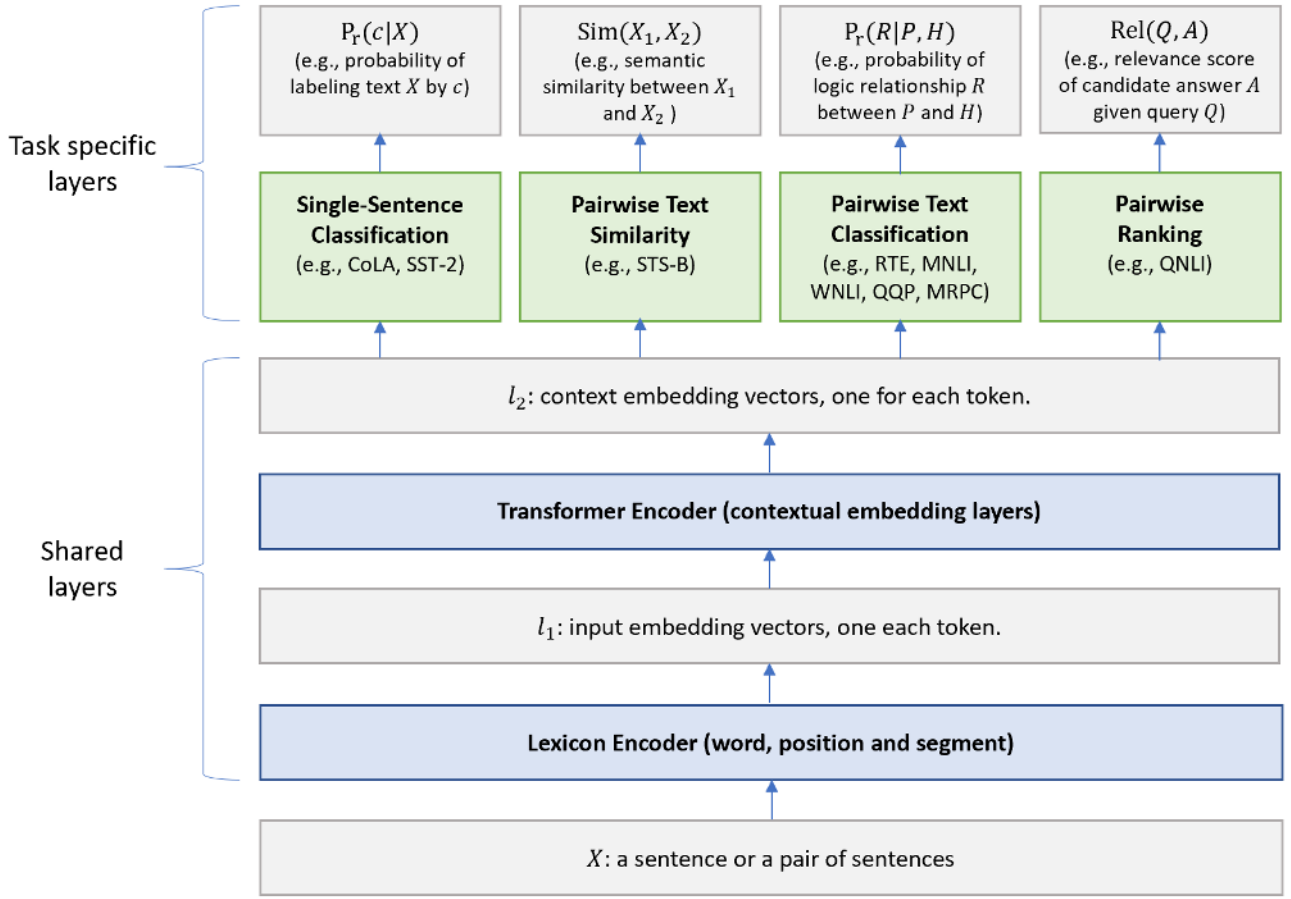
	MUC			B^3			$CEAF_{\phi_4}$			
	P	R	F1	P	R	F1	P	R	F1	Avg. F1
Prev. SotA: (Lee et al., 2018)	81.4	79.5	80.4	72.2	69.5	70.8	68.2	67.1	67.6	73.0
Google BERT	84.9	82.5	83.7	76.7	74.2	75.4	74.6	70.1	72.3	77.1
Our BERT	85.1	83.5	84.3	77.3	75.5	76.4	75.0	71.9	73.9	78.3
Our BERT-1seq	85.5	84.1	84.8	77.8	76.7	77.2	75.3	73.5	74.4	78.8
SpanBERT	**85.8**	**84.8**	**85.3**	**78.3**	**77.9**	**78.1**	**76.4**	**74.2**	**75.3**	**79.6**

GLUE

	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE	(Avg)
Google BERT	59.3	**95.2**	88.5/84.3	86.4/88.0	71.2/89.0	86.1/85.7	93.0	71.1	80.4
Our BERT	58.6	93.9	90.1/86.6	88.4/89.1	71.8/89.3	87.2/86.6	93.0	74.7	81.1
Our BERT-1seq	63.5	94.8	**91.2**/87.8	89.0/88.4	**72.1/89.5**	88.0/87.4	93.0	72.1	81.7
SpanBERT	**64.3**	94.8	90.9/**87.9**	**89.9/89.1**	71.9/**89.5**	**88.1/87.7**	**94.3**	**79.0**	**82.8**

MT-DNN (BigBird) [Liu+ 2019]

- BERT + multi-task training



MT-DNN (BigBird) [Liu+ 2019]

Model	CoLA 8.5k	SST-2 67k	MRPC 3.7k	STS-B 7k	QQP 364k	MNLI-m/mm 393k	QNLI 108k	RTE 2.5k	WNLI 634	AX	**Score**
BiLSTM+ELMo+Attn [1]	36.0	90.4	84.9/77.9	75.1/73.3	64.8/84.7	76.4/76.1	-	56.8	65.1	26.5	70.5
Singletask Pretrain Transformer [2]	45.4	91.3	82.3/75.7	82.0/80.0	70.3/88.5	82.1/81.4	-	56.0	53.4	29.8	72.8
GPT on STILTs [3]	47.2	93.1	87.7/83.7	85.3/84.8	70.1/88.1	80.8/80.6	-	69.1	65.1	29.4	76.9
$\text{BERT}_{\text{LARGE}}$ [4]	60.5	94.9	89.3/85.4	87.6/86.5	72.1/89.3	86.7/85.9	92.7	70.1	65.1	39.6	80.5
$\text{MT-DNN}_{\text{no-fine-tune}}$	58.9	94.6	**90.1/86.4**	89.5/88.8	**72.7/89.6**	86.5/85.8	**93.1**	79.1	65.1	39.4	81.7
MT-DNN	**62.5**	**95.6**	**91.1/88.2**	**89.5/88.8**	**72.7/89.6**	**86.7/86.0**	**93.1**	**81.4**	65.1	**40.3**	**82.7**
Human Performance	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	-	87.1

ERNIE 2.0 [Sun+ 2019] (Baidu)

a Continual Pre-training Framework for Language Understanding

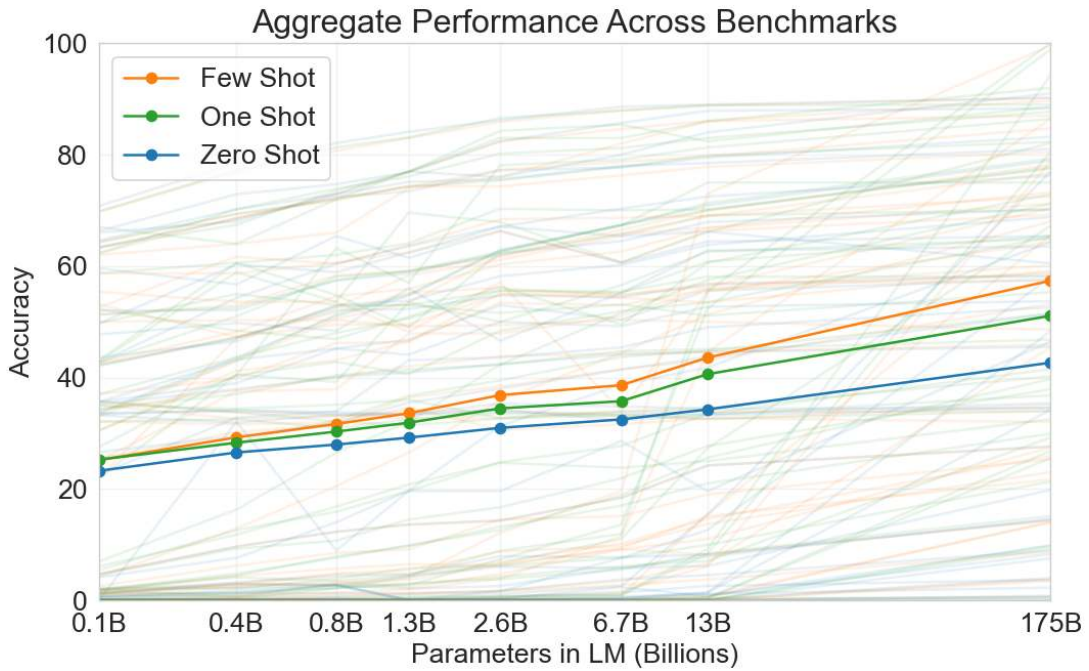
- Pre-training tasks are incrementally added
  - Task A → Tasks A + B → Tasks A + B + C → ...
- Pre-training tasks:
  - Word-aware tasks
    - Knowledge masking (phrase/entity masking; ERNIE 1.0)
    - Capitalization prediction
    - Token-document relation prediction
  - Structure-aware tasks
    - Sentence reordering
    - Sentence distance
  - Semantic-aware tasks
    - Discourse relation
    - IR relevance

ERNIE 2.0 [Sun+ 2019]

Task(Metrics)	*BASE model*		*LARGE model*				
	Test		Dev			Test	
	BERT	ERNIE 2.0	BERT	XLNet	ERNIE 2.0	BERT	ERNIE 2.0
CoLA (Matthew Corr.)	52.1	**55.2**	60.6	63.6	**65.4**	60.5	**63.5**
SST-2 (Accuracy)	93.5	**95.0**	93.2	95.6	**96.0**	94.9	**95.6**
MRPC (Accurary/F1)	84.8/88.9	**86.1/89.9**	88.0/-	89.2/-	**89.7/-**	85.4/89.3	**87.4/90.2**
STS-B (Pearson Corr./Spearman Corr.)	87.1/85.8	**87.6/86.5**	90.0/-	91.8/-	**92.3/-**	87.6/86.5	**91.2/90.6**
QQP (Accuracy/F1)	89.2/71.2	**89.8/73.2**	91.3/-	91.8/-	**92.5/-**	89.3/72.1	**90.1/73.8**
MNLI-m/mm (Accuracy)	84.6/83.4	**86.1/85.5**	86.6/-	**89.8/-**	89.1/-	86.7/85.9	**88.7/88.8**
QNLI (Accuracy)	90.5	**92.9**	92.3	93.9	**94.3**	92.7	**94.6**
RTE (Accuracy)	66.4	**74.8**	70.4	83.8	**85.2**	70.1	**80.2**
WNLI (Accuracy)	**65.1**	**65.1**	-	-	-	65.1	**67.8**
AX(Matthew Corr.)	34.2	**37.4**	-	-	-	39.6	**48.0**
Score	78.3	**80.6**	-	-	-	80.5	**83.6**

Language Models

- GPT-2: zero-shot
- GPT-3: few-shot



The three settings we explore for in-context learning

Zero-shot

The model predicts the answer given only a natural language discription of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```

Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

```
1  sea otter => loutre de mer          ← example #1
```

gradient update

```
1  peppermint => menthe poivrée        ← example #2
```

gradient update

...

```
1  plush giraffe => girafe peluche     ← example #N
```

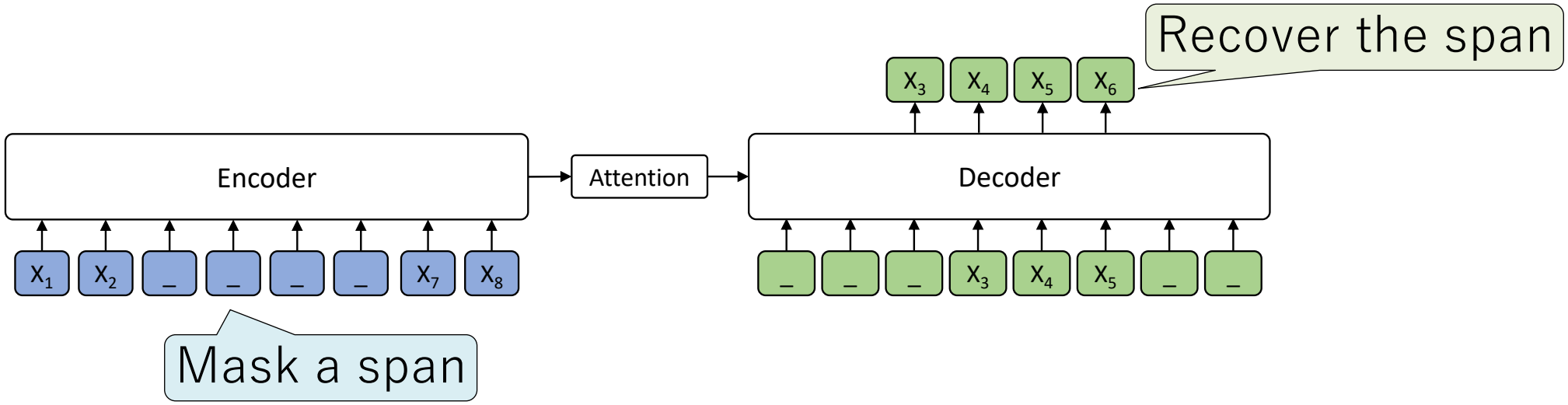
gradient update

```
1  cheese =>                           ← prompt
```

[Brown+ 2020]

Generation Models based on Pre-training

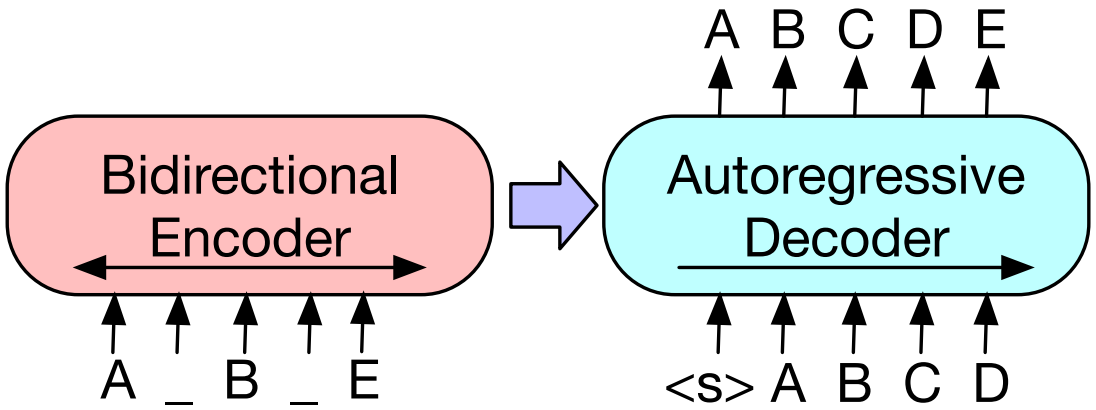
- MASS (Masked Sequence to Sequence Pre-training) [Song+ 19]



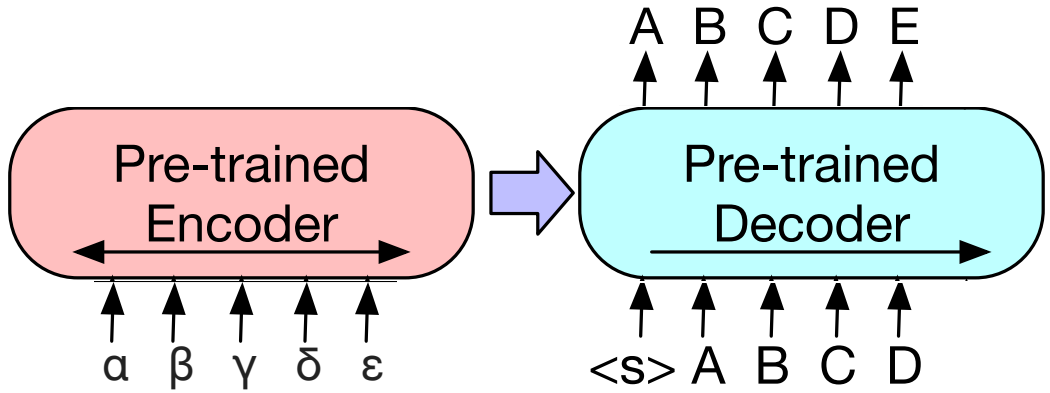
MASS was effective mainly for unsupervised machine translation

BART
(Bidirectional and Auto-Regressive Transformers)

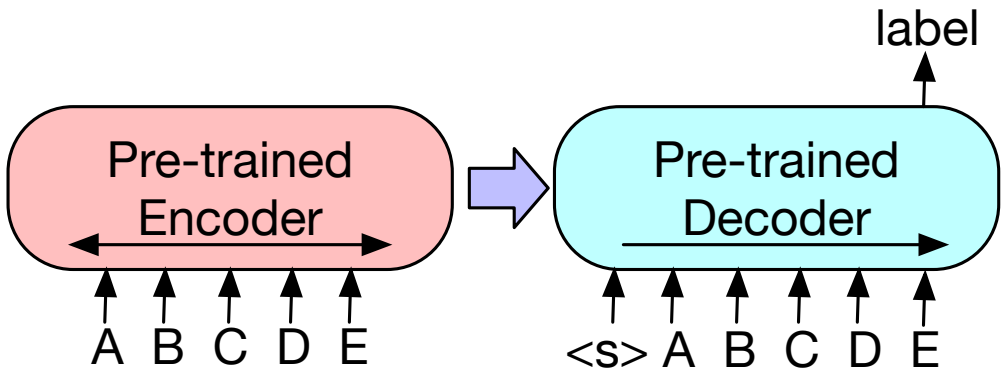
[Lewis+ 19]



Pre-training



Fine-tuning for generation tasks



Fine-tuning for classification tasks

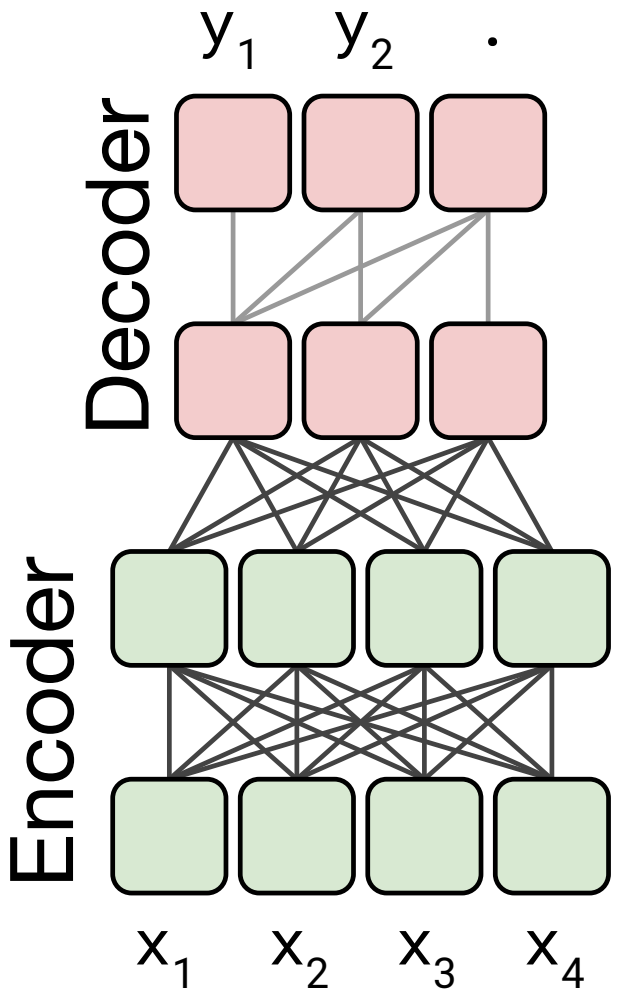
Text-to-Text Transfer Transformer (T5) [Raffel+ 2019]

- All tasks are converted to a text-to-text format

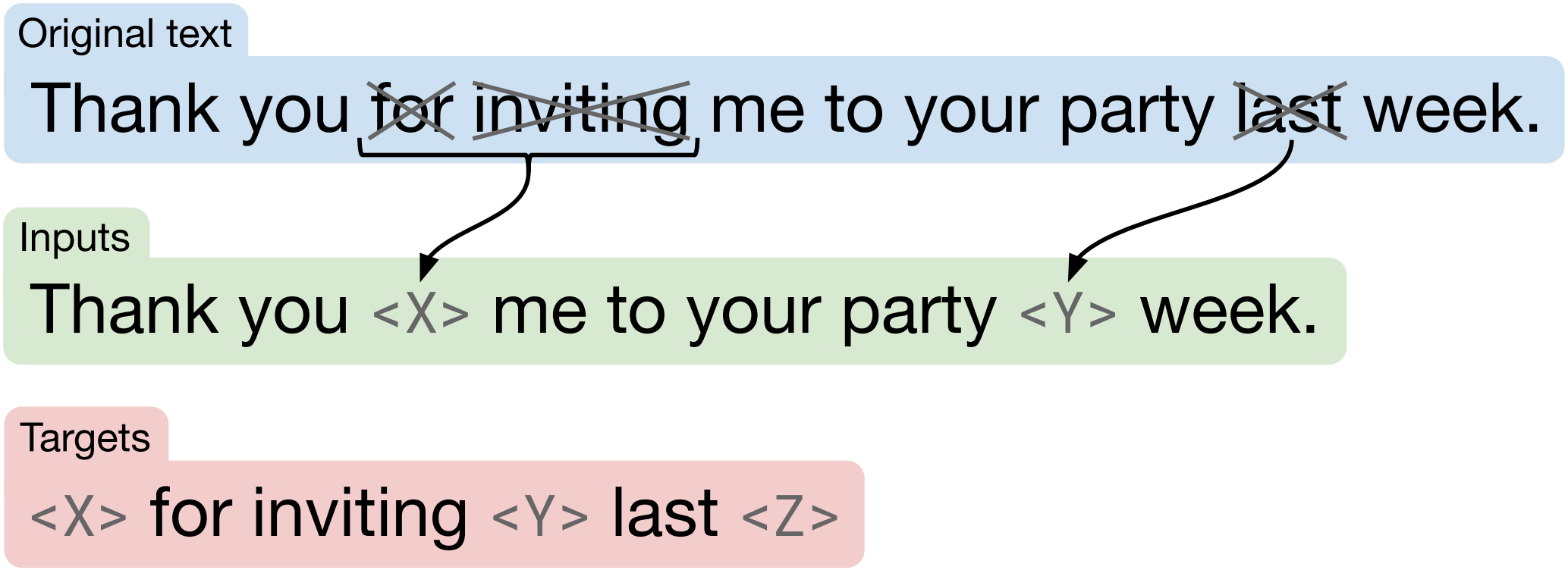
task	source	target
COLA	cola sentence: The course is jumping well.	not acceptable
STS-B	stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field.	3.8
Summarization	summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in Mississippi ...	six people hospitalized after a storm in attala county.
MT	translate English to German: That is good.	Das ist gut.
WSC	wsc: The city councilmen refused the demonstrators a permit because *they* feared violence.	The city councilmen

T5: Model

- An encoder-decoder model
  - Each consists of Transformer



T5: Pre-training



T5: Results

Model	GLUE Average	CoLA Matthew's	SST-2 Accuracy	MRPC F1	MRPC Accuracy	STS-B Pearson	STS-B Spearman
Previous best	89.4[a]	69.2[b]	**97.1**[a]	**93.6**[b]	**91.5**[b]	**92.7**[b]	**92.3**[b]
T5-Small	77.4	41.0	91.8	89.7	86.6	85.6	85.0
T5-Base	82.7	51.1	95.2	90.7	87.5	89.4	88.6
T5-Large	86.4	61.2	96.3	92.4	89.9	89.9	89.2
T5-3B	88.5	67.1	97.4	92.5	90.0	90.6	89.8
T5-11B	**89.7**	**70.8**	**97.1**	91.9	89.2	92.5	92.1

Model	QQP F1	QQP Accuracy	MNLI-m Accuracy	MNLI-mm Accuracy	QNLI Accuracy	RTE Accuracy	WNLI Accuracy
Previous best	**74.8**[c]	**90.7**[b]	91.3[a]	91.0[a]	**99.2**[a]	89.2[a]	91.8[a]
T5-Small	70.0	88.0	82.4	82.3	90.3	69.9	69.2
T5-Base	72.6	89.4	87.1	86.2	93.7	80.1	78.8
T5-Large	73.9	89.9	89.9	89.6	94.8	87.2	85.6
T5-3B	74.4	89.7	91.4	91.2	96.3	91.1	89.7
T5-11B	74.6	90.4	**92.0**	**91.7**	96.7	**92.5**	**93.2**

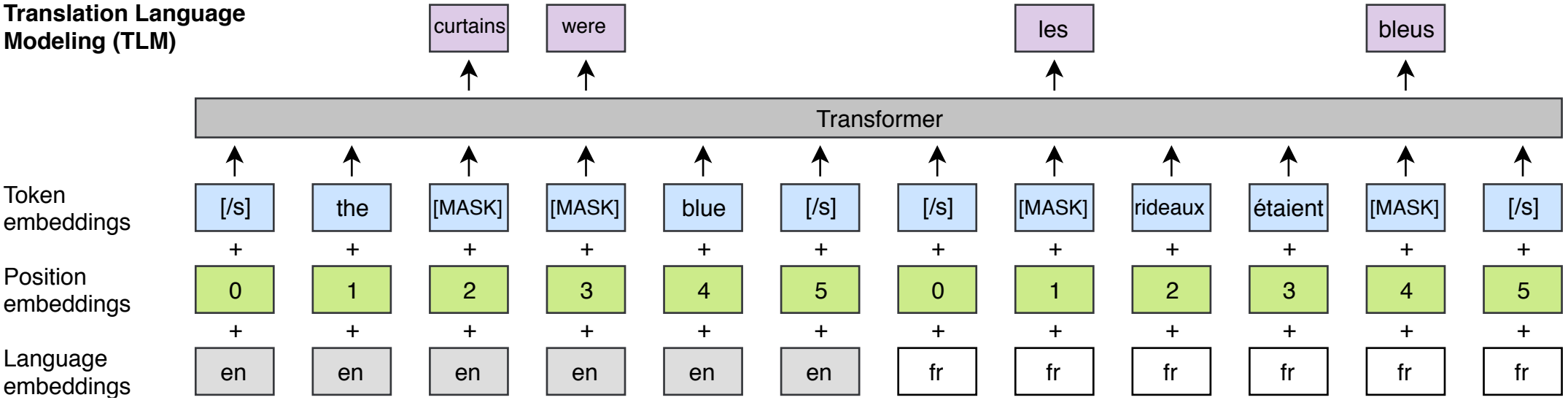
Model	SQuAD EM	SQuAD F1	SuperGLUE Average	BoolQ Accuracy	CB F1	CB Accuracy	COPA Accuracy
Previous best	88.95[d]	94.52[d]	84.6[e]	87.1[e]	90.5[e]	95.2[e]	90.6[e]
T5-Small	79.10	87.24	63.3	76.4	56.9	81.6	46.0
T5-Base	85.44	92.08	76.2	81.4	86.2	94.0	71.2
T5-Large	86.66	93.79	82.3	85.4	91.6	94.8	83.4
T5-3B	88.53	94.95	86.4	89.9	90.3	94.4	92.0
T5-11B	**90.06**	**95.64**	**88.9**	**91.0**	**93.0**	**96.4**	**94.8**

Model	MultiRC F1a	MultiRC EM	ReCoRD F1	ReCoRD Accuracy	RTE Accuracy	WiC Accuracy	WSC Accuracy
Previous best	84.4[e]	52.5[e]	90.6[e]	90.0[e]	88.2[e]	69.9[e]	89.0[e]
T5-Small	69.3	26.3	56.3	55.4	73.3	66.9	70.5
T5-Base	79.7	43.1	75.0	74.2	81.5	68.3	80.8
T5-Large	83.3	50.7	86.8	85.9	87.8	69.3	86.3
T5-3B	86.8	58.3	91.2	90.4	90.7	72.1	90.4
T5-11B	**88.2**	**62.3**	**93.3**	**92.5**	**92.5**	**76.1**	**93.8**

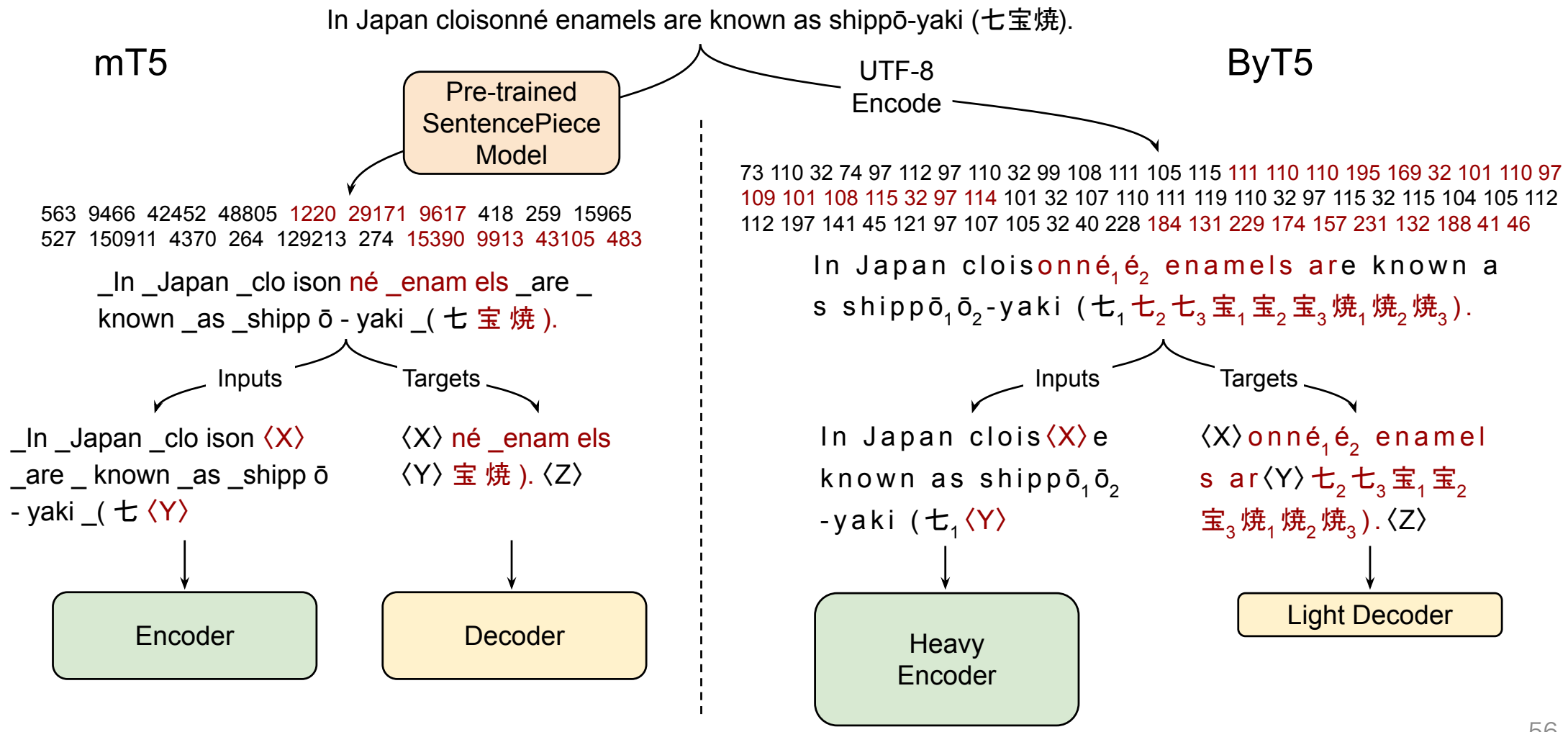
Model	WMT EnDe BLEU	WMT EnFr BLEU	WMT EnRo BLEU	CNN/DM ROUGE-1	CNN/DM ROUGE-2	CNN/DM ROUGE-L
Previous best	**33.8**[f]	**43.8**[f]	**38.5**[g]	43.47[h]	20.30[h]	40.63[h]
T5-Small	26.7	36.0	26.8	41.12	19.56	38.35
T5-Base	30.9	41.2	28.0	42.05	20.34	39.40
T5-Large	32.0	41.5	28.1	42.50	20.68	39.75
T5-3B	31.8	42.6	28.2	42.72	21.02	39.94
T5-11B	32.1	43.4	28.1	**43.52**	**21.55**	**40.69**

Multilingualization

- Pre-training on multilingual corpora (25-100 languages)
  - XLM [Lample+ 19], XLM-R [Conneau+ 19], mBART [Liu+ 20], mT5 [Xue+ 20], ByT5 [Xue+ 21]
- Pre-training on bilingual corpora (XLM [Lample+ 19])



Token-free Model: ByT5 [Xue+ 21]

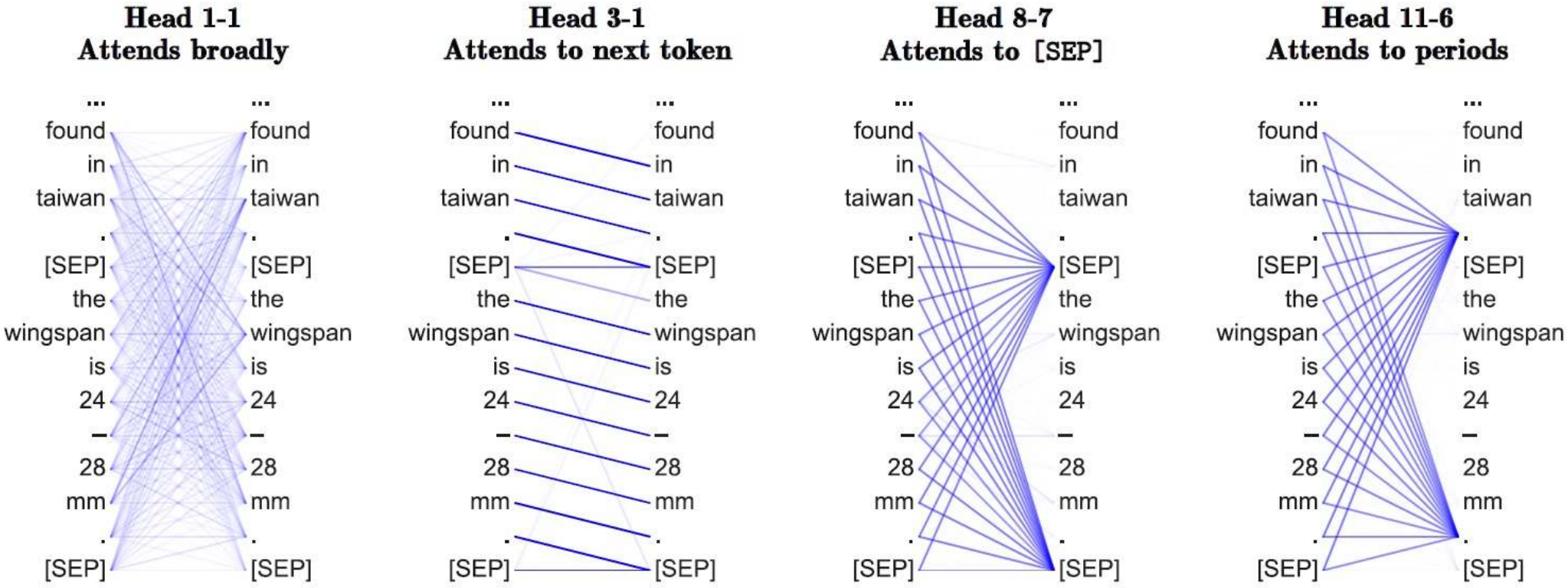


Token-free Model: ByT5 [Xue+ 21]

	Small		Base		Large		XL		XXL	
	mT5	ByT5	mT5	ByT5	mT5	ByT5	mT5	ByT5	mT5	ByT5
In-language multitask (models fine-tuned on gold data in all target languages)										
WikiAnn NER	86.4	**90.6**	88.2	**91.6**	89.7	**91.8**	91.3	**92.6**	92.2	**93.7**
TyDiQA-GoldP	74.0 / 62.7	**82.6 / 73.6**	79.7 / 68.4	**86.4 / 78.0**	85.3 / 75.3	**87.7 / 79.2**	87.6 / 78.4	**88.0 / 79.3**	88.7 / 79.5	**89.4 / 81.4**
Translate-train (models fine-tuned on English data plus translations in all target languages)										
XNLI	72.0	**76.6**	79.8	**79.9**	**84.4**	82.8	**85.3**	85.0	**87.1**	85.7
PAWS-X	79.9	**88.6**	89.3	**89.8**	**91.2**	90.6	**91.0**	90.5	91.5	**91.7**
XQuAD	64.3 / 49.5	**74.0 / 59.9**	75.3 / 59.7	**78.5 / 64.6**	81.2 / 65.9	**81.4 / 67.4**	82.7 / 68.1	**83.7 / 69.5**	**85.2 / 71.3**	84.1 / 70.2
MLQA	56.6 / 38.8	**67.5 / 49.9**	67.6 / 48.5	**71.9 / 54.1**	73.9 / 55.2	**74.4 / 56.1**	75.1 / 56.6	**75.9 / 57.7**	**76.9** / 58.3	**76.9 / 58.8**
TyDiQA-GoldP	49.8 / 35.6	**64.2 / 50.6**	66.4 / 51.0	**75.6 / 61.7**	75.7 / 60.1	**80.1 / 66.4**	80.1 / 65.0	**81.5 / 67.6**	**83.3** / 69.4	83.2 / **69.6**
Cross-lingual zero-shot transfer (models fine-tuned on English data only)										
XNLI	67.5	**69.1**	**75.4**	**75.4**	**81.1**	79.7	**82.9**	82.2	**85.0**	83.7
PAWS-X	82.4	**84.0**	**86.4**	86.3	**88.9**	87.4	**89.6**	88.6	90.0	**90.1**
WikiAnn NER	50.5	**57.6**	55.7	**62.0**	58.5	**62.9**	**65.5**	61.6	**69.2**	67.7
XQuAD	58.1 / 42.5	**66.3 / 49.7**	**67.0 / 49.0**	66.6 / 48.1	**77.8 / 61.5**	61.5 / 43.9	**79.5 / 63.6**	57.7 / 43.0	**82.5 / 66.8**	79.7 / 63.6
MLQA	54.6 / 37.1	**60.9 / 43.3**	64.4 / 45.0	**66.6 / 47.3**	**71.2 / 51.7**	65.6 / 45.0	**73.5 / 54.4**	65.1 / 46.5	**76.0 / 57.4**	71.6 / 54.9
TyDiQA-GoldP	36.4 / 24.4	**54.9 / 39.9**	59.1 / 42.4	**69.6 / 54.2**	68.4 / 50.9	**75.4 / 59.4**	**77.8 / 61.8**	63.2 / 49.2	**82.0 / 67.3**	75.3 / 60.0

Analysis of Inside BERT

Analysis of attention [Clark+ 19]

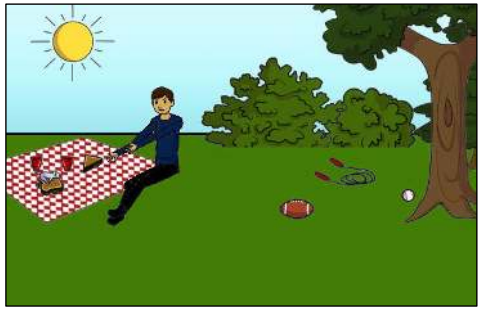


Multimodal Tasks of Language and Vision

- Visual Question Answering (VQA) [Agrawal+ 15]

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

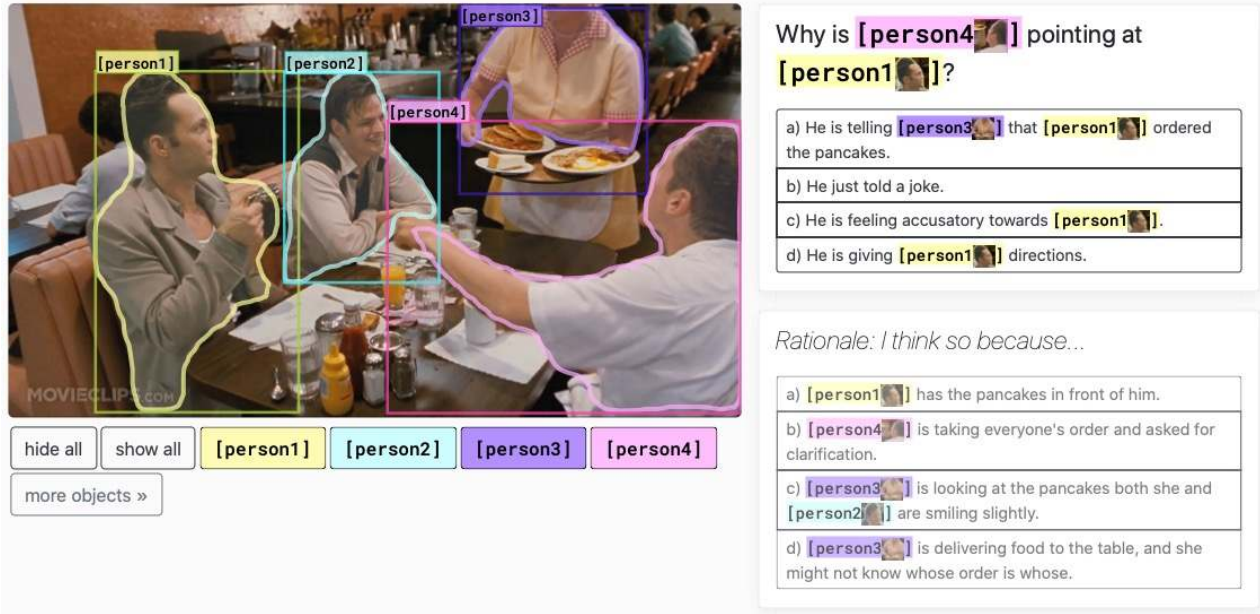


Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

https://visualqa.org/

- Visual Commonsense Reasoning (VCR) [Zellers+ 18]



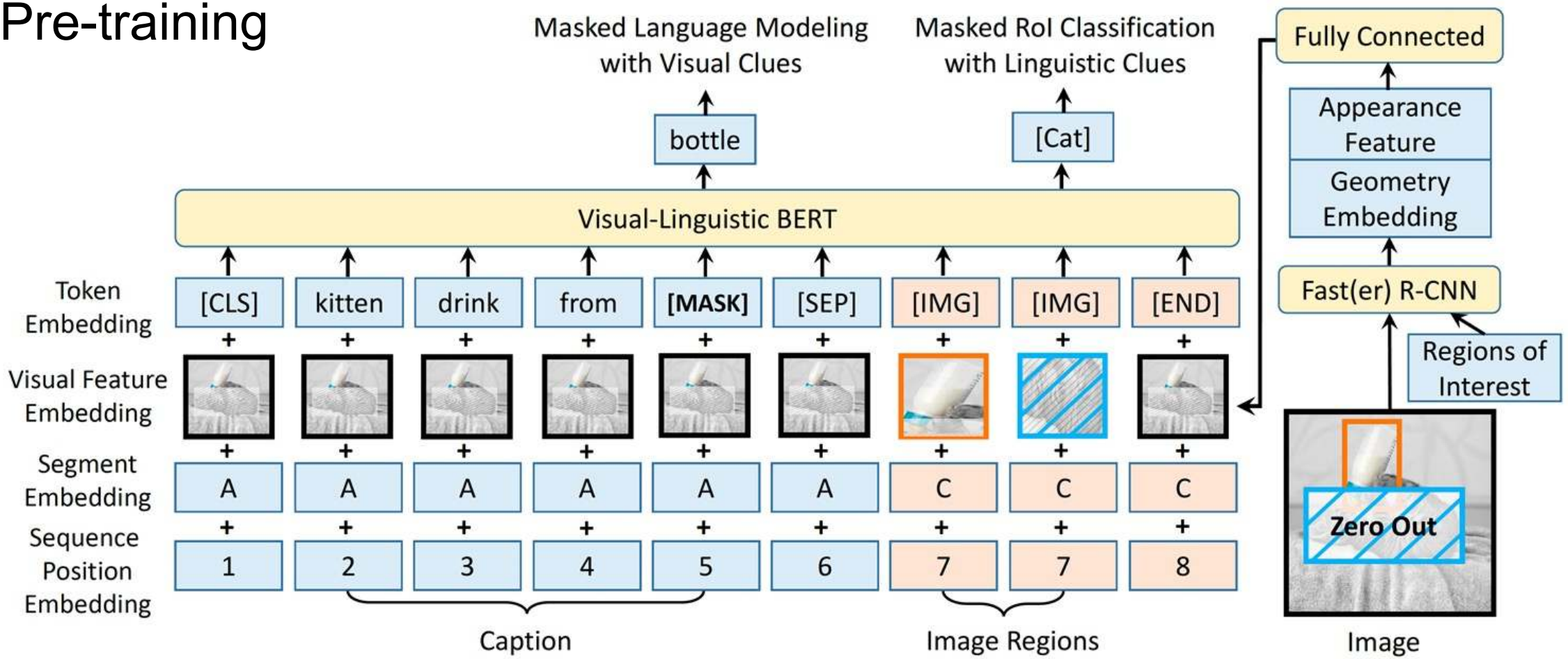
https://visualcommonsense.com/

Models for Multimodal Tasks

- Input: 1 sequence
  - Encode a sequence that concatenates text and images
    - VideoBERT [Sun+ 19], VisualBERT [Li+ 19], Unicoder-VL [Li+ 19], VL-BERT [Su+ 19], UNITER [Chen+ 19], ImageBERT [Qi+ 20]
- Input: 2 sequences
  - Encode text and images respectively and then fuse them
    - CBT [Sun+ 19], ViLBERT [Lu+ 19], LXMERT [Tan+ 19], ERNIE-ViL [Yu+ 20]

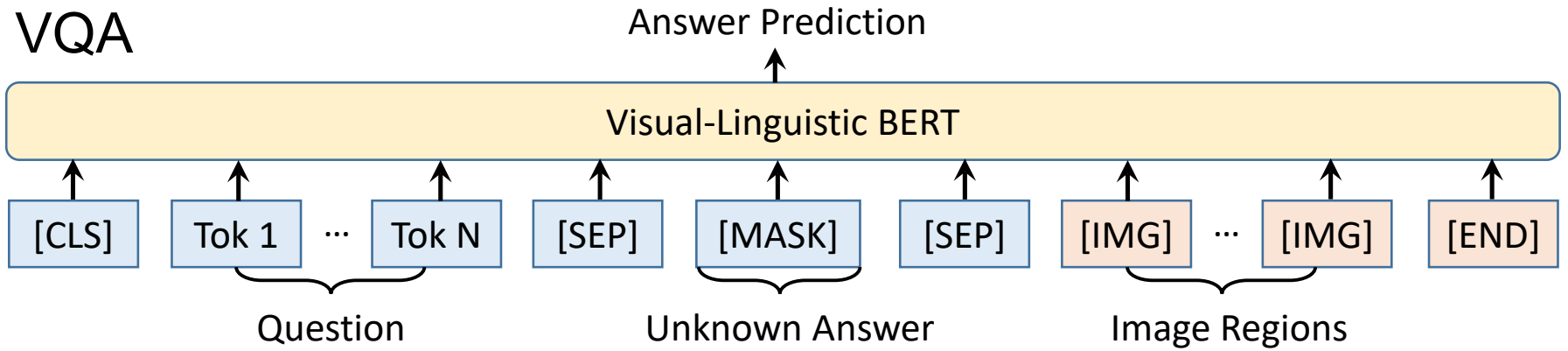
1-sequence Model: VL-BERT [Su+ 19]

- Pre-training

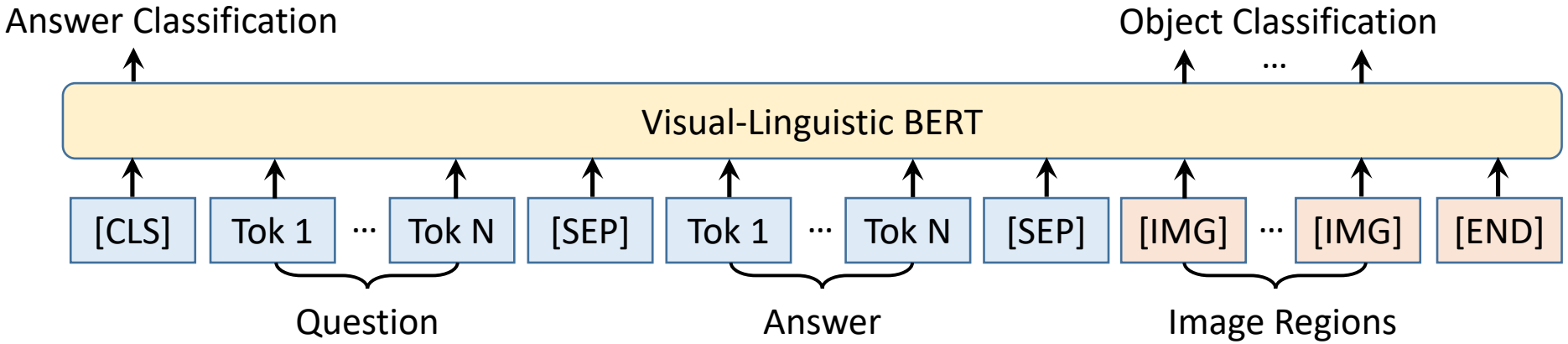


1-sequence Model: VL-BERT [Su+ 19]

- Fine-tuning
  - VQA

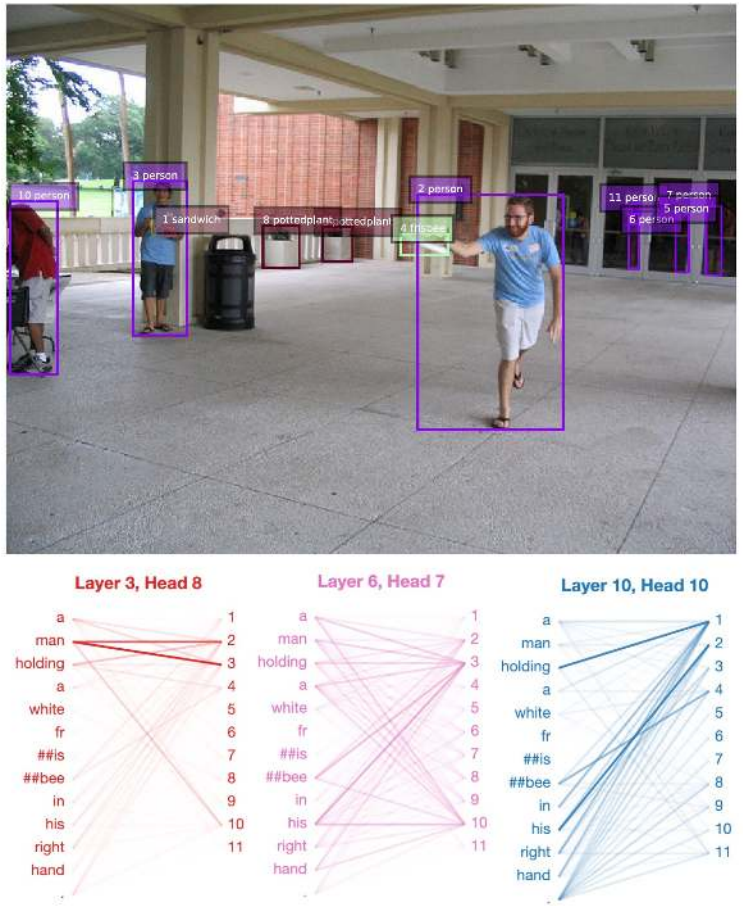


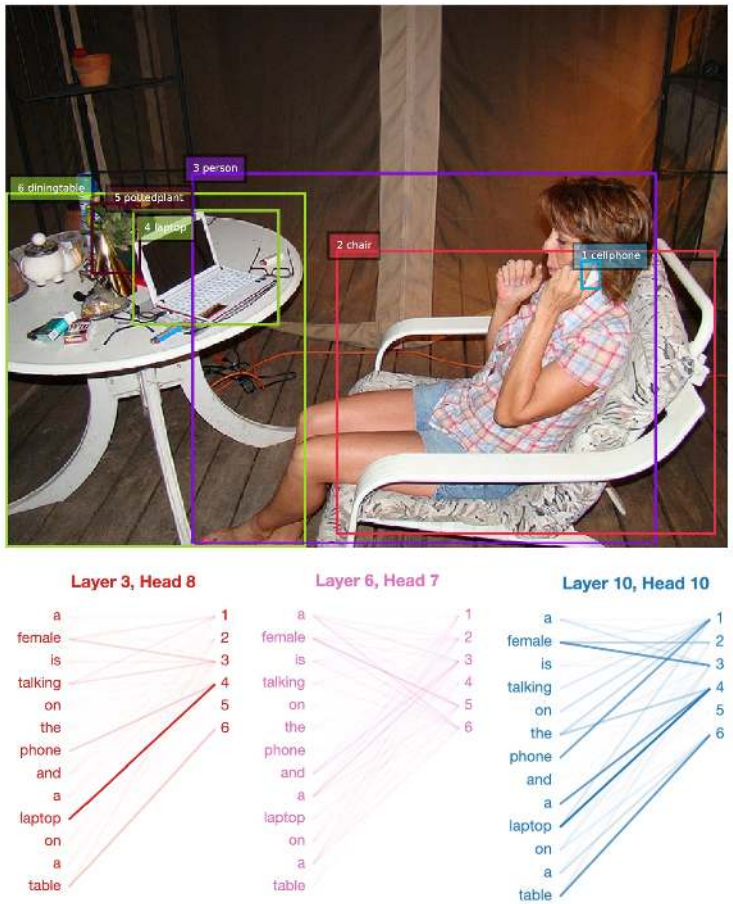
  - VCR



1-sequence Model: VL-BERT [Su+ 19]

- Visualization





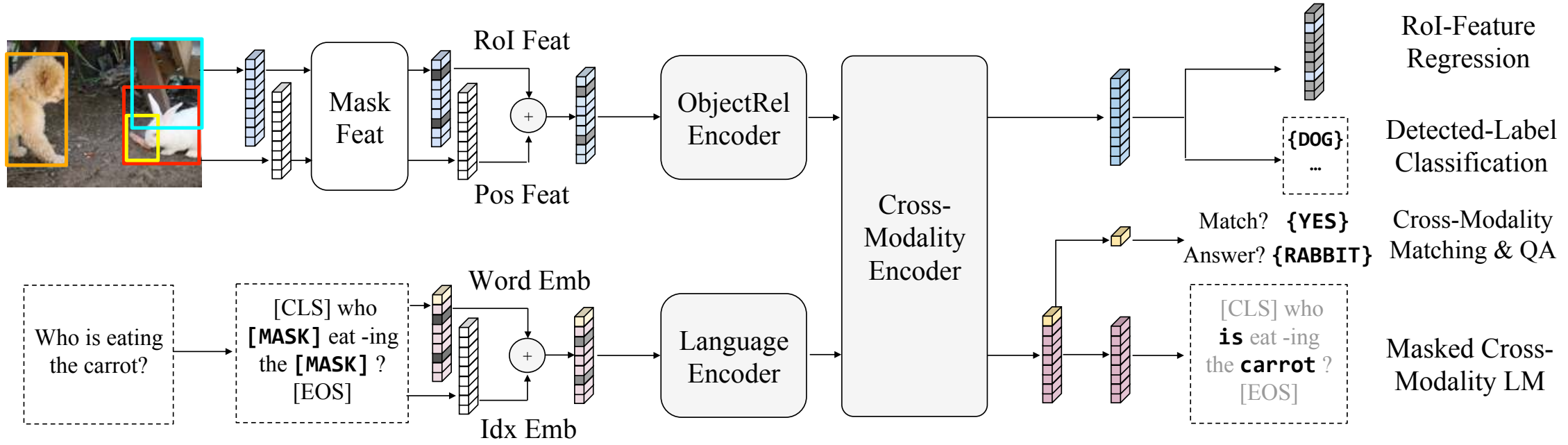
2-sequence Model: LXMERT

[Tan+ 19]

Model	test-dev	test-std
BUTD (Anderson et al., 2018)	65.32	65.67
ViLBERT (Lu et al., 2019)†	70.55	70.92
VisualBERT (Li et al., 2019b)†	70.80	71.00
LXMERT (Tan & Bansal, 2019)†	72.42	72.54
VL-BERT$_{BASE}$ w/o pre-training	69.58	-
VL-BERT$_{BASE}$	71.16	-
VL-BERT$_{LARGE}$	71.79	72.22

Performance on VQA (from [Su+ 19])

- Pre-training



Libraries

- Pre-training
  - TensorFlow (by Google)
    - https://github.com/google-research/bert (TF 1.0)
    - https://github.com/tensorflow/models/tree/master/official/nlp/bert (TF 2.0)
      - Multi-GPU is supported
- Fine-tuning
  - TensorFlow (by Google)
  - PyTorch: Transformers (by Hugging Face)
    - https://github.com/huggingface/transformers
    - Latest models, such as RoBERTa and T5, are also supported