Recent Progress in Pre-training for NLP

Furu Wei

Senior Principal Research Manager (首席研究员), MSRA\NLP

Presented at CCMT 2020

Joint work with my colleagues and interns at MSRA

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Transformer (NeurIPS 2017)

NMT

BERT (NAACL 2019)

"NMT"

The Era of NLP & Pre-training

OpenAI GPT-3

May 2020

175 billion parameters

*The next stage of an artificial intelligence that was dubbed '**too dangerous to release**'*

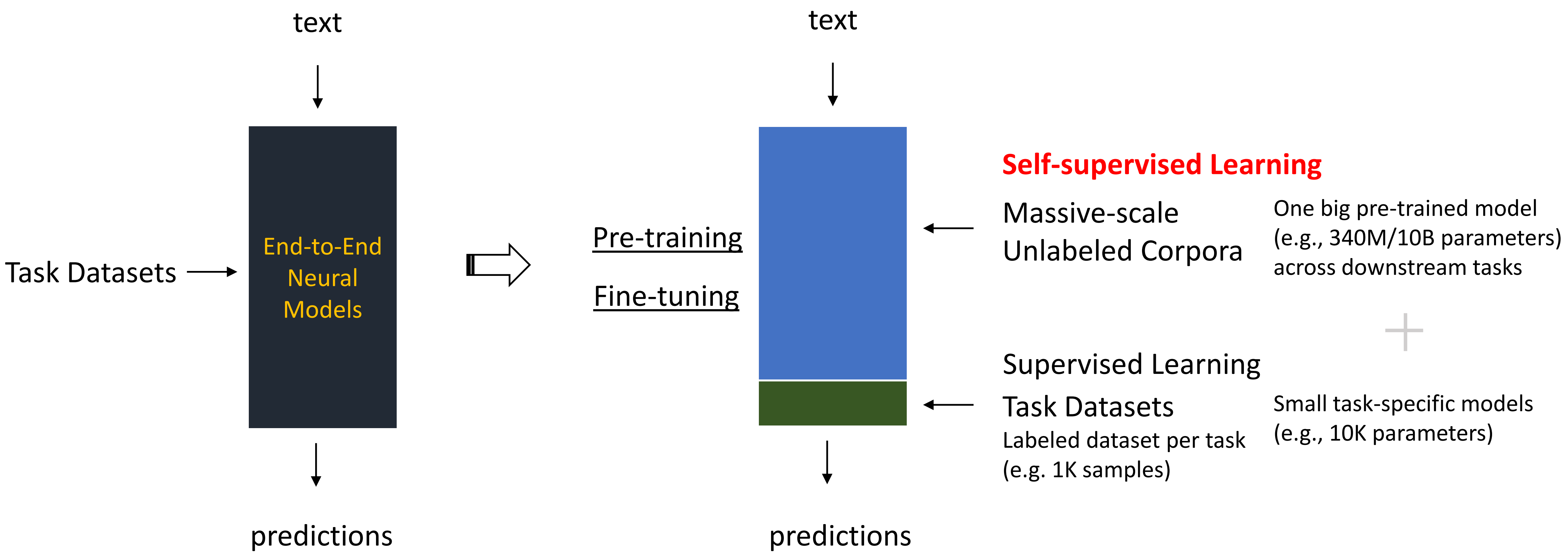
Google AI BERT

Oct. 2018

11 language tasks

***Fundamentally** changes how we build SOTA NLP (and NL related) models*

New NLP Paradigm: Pre-training + Fine-tuning



Downstream NLP Tasks: text classification, entity recognition, question answering, summarization, …

Pre-trained Model Task-specific Model (Layer)

Transfer Learning for NLP Tasks

Language skills

Classify text into topics or sentiments

Extract entities from text

Find answers to questions from text

Summarize long text (e.g. a document)

…

Understand the **meaning** of text (e.g. words and their relations) in **context**

Reading (Cloze test)

NLP tasks

Text classification

Entity recognition

Question answering

Text summarization

…

Contextualized representations (vectors)

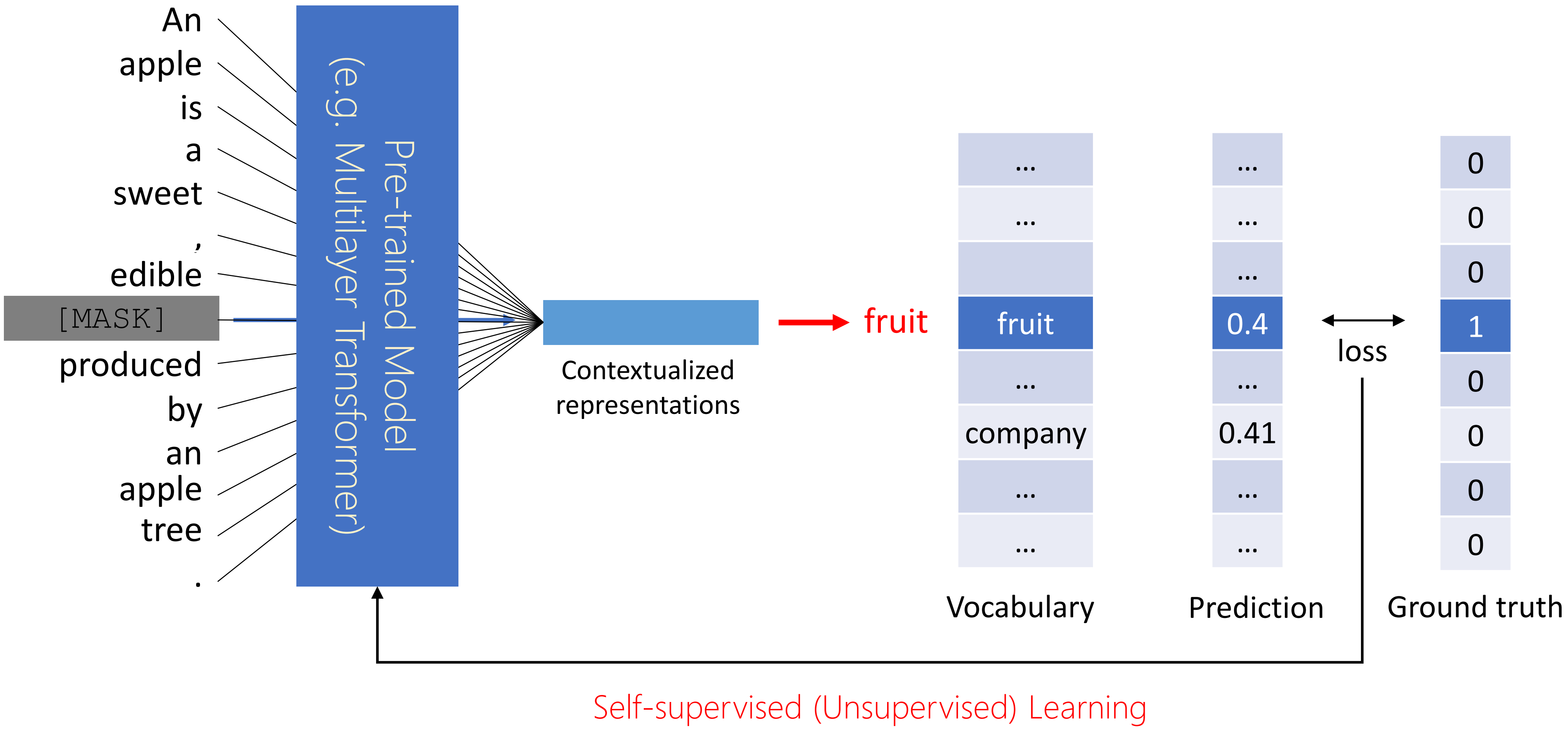
Self-supervised learning (Masked LM)

Massive-scale self-supervised learning via cloze tests

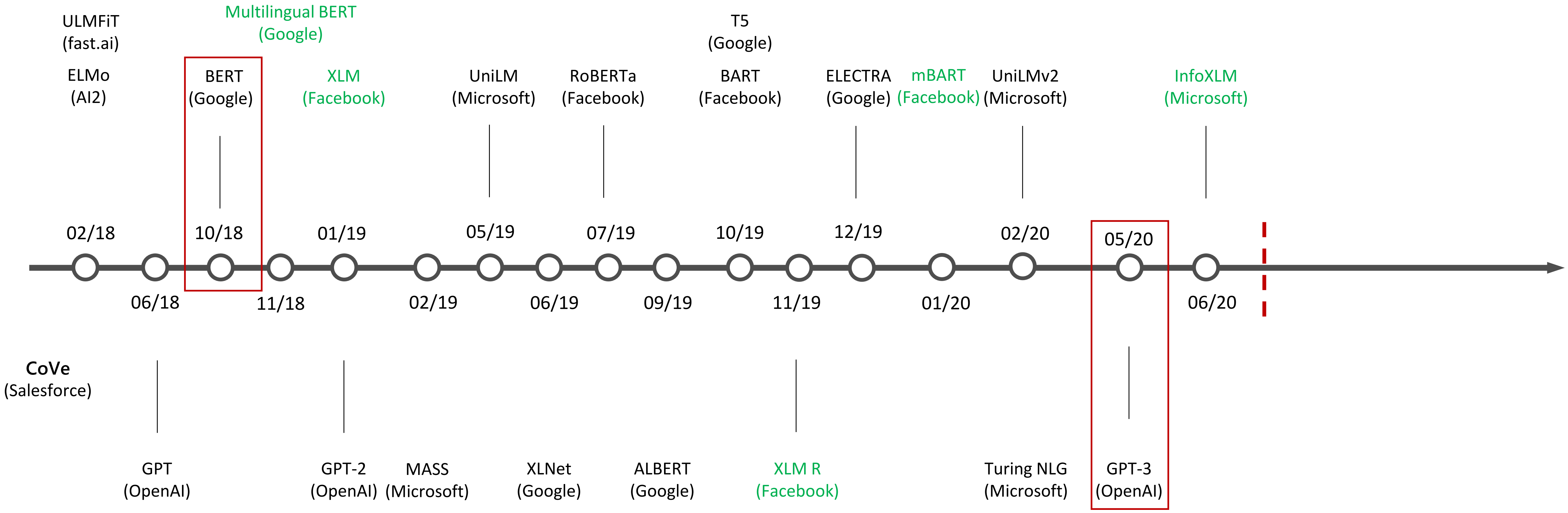
fruit ?
company ?
…

Language Modeling (LM) / Cloze tests as pre-training tasks/objectives

An apple is a sweet, edible ____ produced by an apple tree.



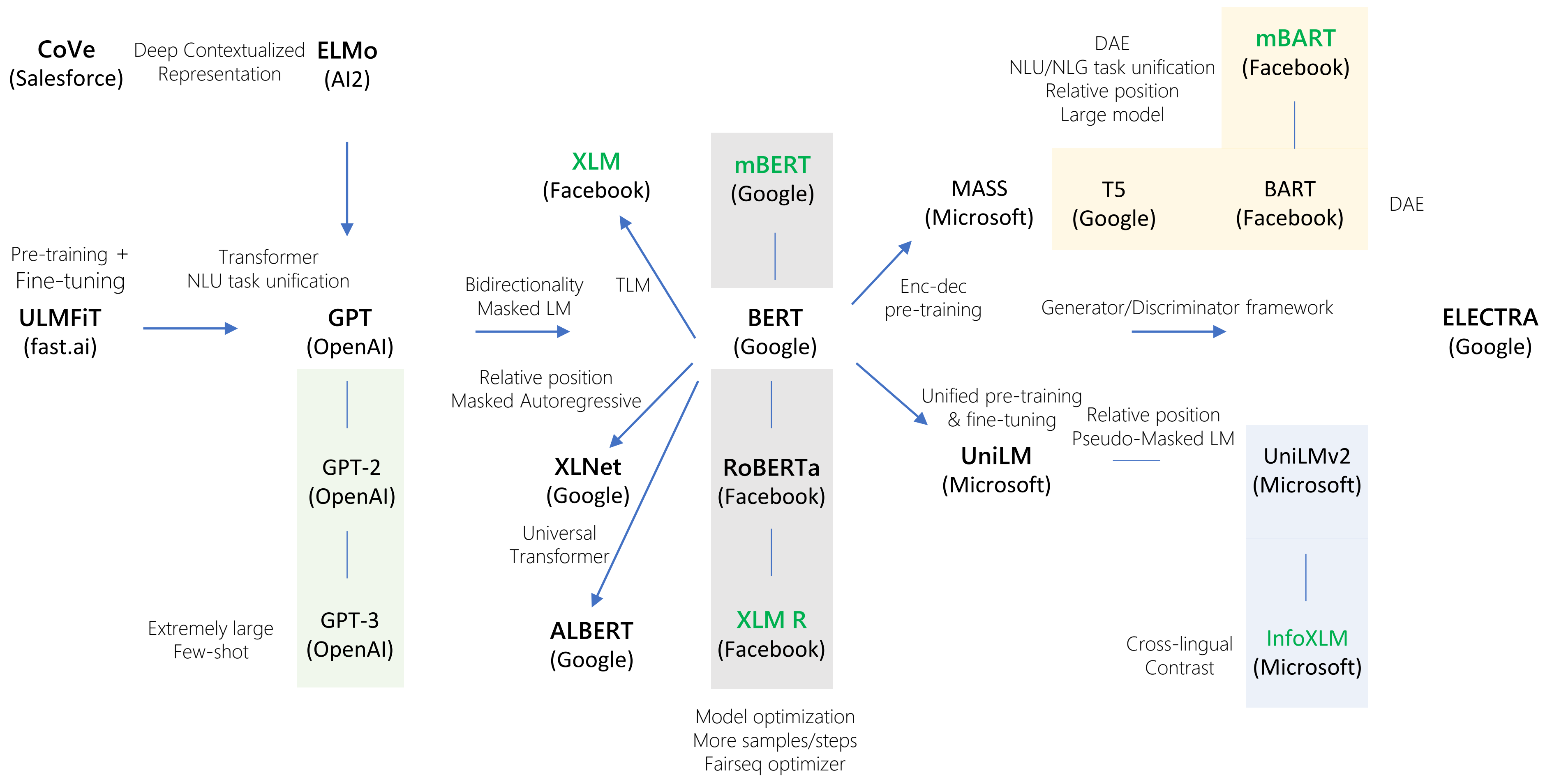
An apple is a sweet, edible fruit produced by an apple tree.



Key trends

- Big model, large corpora
- Unified model for NLU, NLG and multilingual tasks (incl. MT)
- Novel frameworks, pre-training tasks and network structures

The Anatomy of Language Model Pre-training



The Anatomy of Language Model Pre-training

NLP Tasks

Task Layers

Backbone Networks

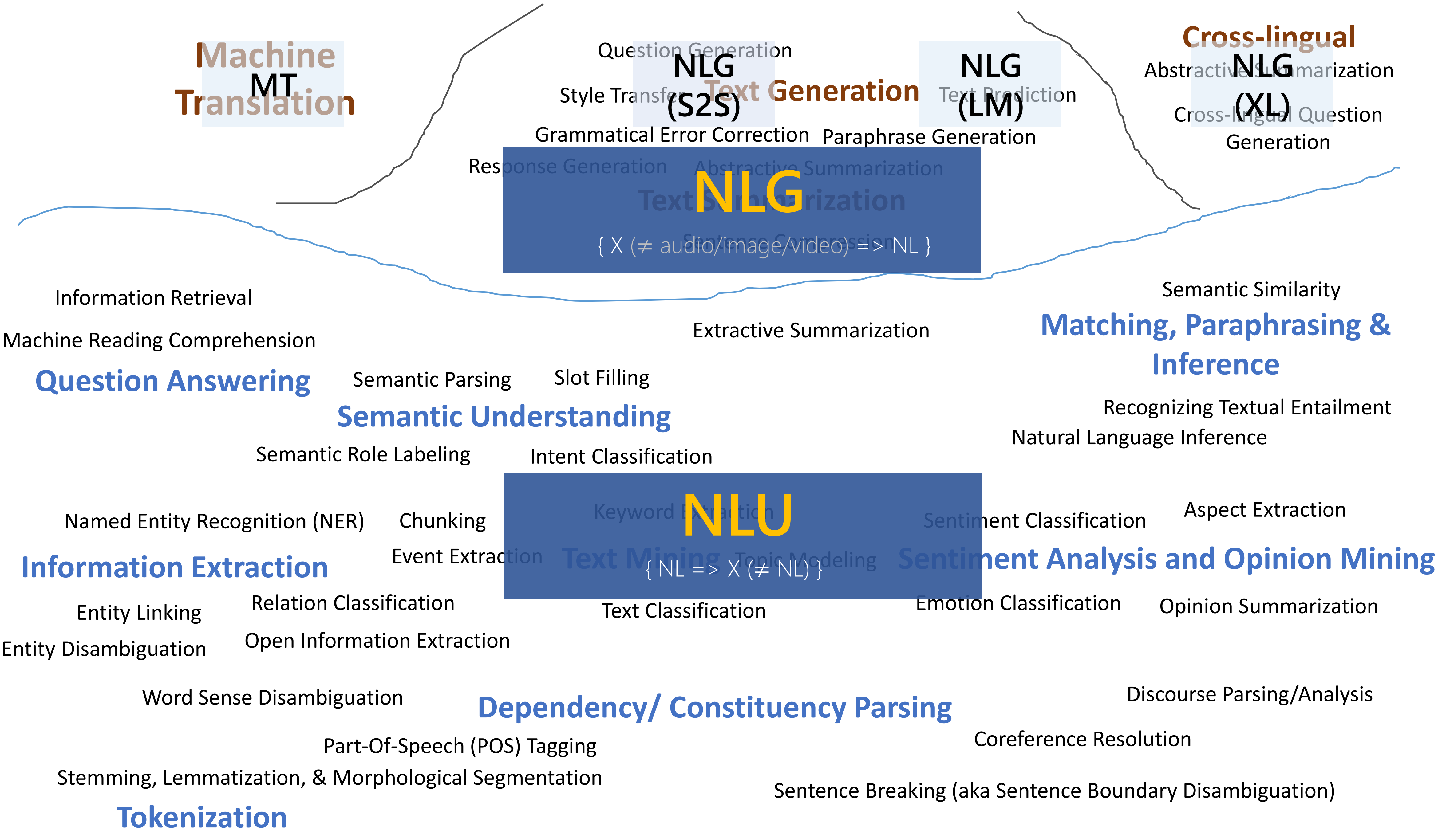
(e.g. Multilayer Transformer)

Architecture: enc/unified, dec, enc-dec

Pre-training Tasks

Self-supervised Tasks

(e.g. Cloze / Masked LM)



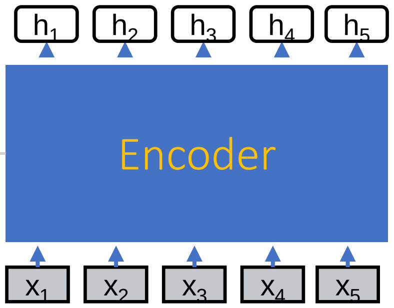
Machine Translation
MT
Question Generation
NLG (S2S)
Style Transfer
Text Generation
NLG (LM)
Text Prediction
Cross-lingual
NLG (XL)
Abstractive Summarization
Cross-lingual Question Generation
Grammatical Error Correction
Paraphrase Generation
Response Generation
Abstractive Summarization
NLG
Text Summarization
Sentence Compression
{ X (≠ audio/image/video) => NL }
Information Retrieval
Machine Reading Comprehension
Question Answering
Extractive Summarization
Semantic Similarity
Matching, Paraphrasing & Inference
Semantic Parsing
Slot Filling
Semantic Understanding
Recognizing Textual Entailment
Natural Language Inference
Semantic Role Labeling
Intent Classification
Named Entity Recognition (NER)
Chunking
Keyword Extraction
NLU
Sentiment Classification
Aspect Extraction
Information Extraction
Event Extraction
Text Mining
Topic Modeling
{ NL => X (≠ NL) }
Sentiment Analysis and Opinion Mining
Entity Linking
Relation Classification
Text Classification
Emotion Classification
Opinion Summarization
Entity Disambiguation
Open Information Extraction
Word Sense Disambiguation
Dependency/ Constituency Parsing
Discourse Parsing/Analysis
Part-Of-Speech (POS) Tagging
Coreference Resolution
Stemming, Lemmatization, & Morphological Segmentation
Sentence Breaking (aka Sentence Boundary Disambiguation)
Tokenization

NLP Tasks & Neural NLP Architectures

NLP Tasks

NLU { NL => X (≠ NL) }

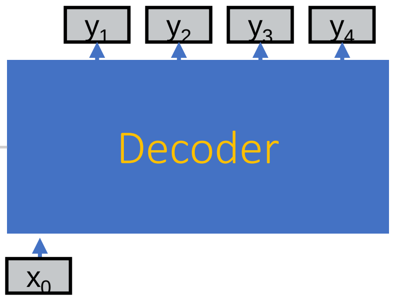
Text (Language E) => { Class, Sequence of class, Score, …}



NLG { X (≠ audio/image/video) => NL }

NLG (LM)

{ } => Text (Language E)

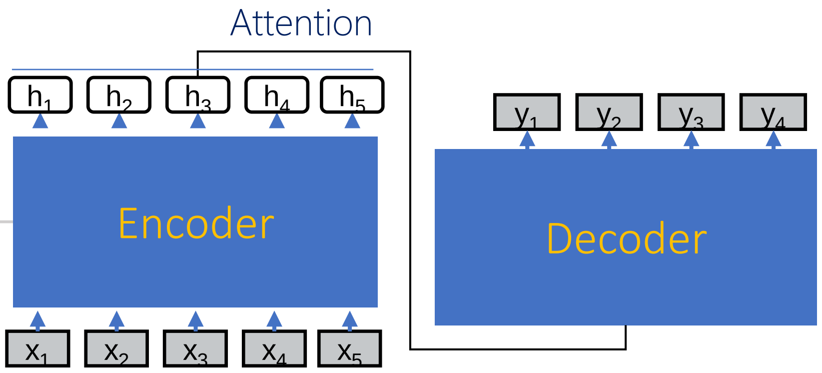


NLG (S2S)

Text (Language E) => Text (Language E)

NLG (XL)/MT

Text (Language E) => Text (Language F)

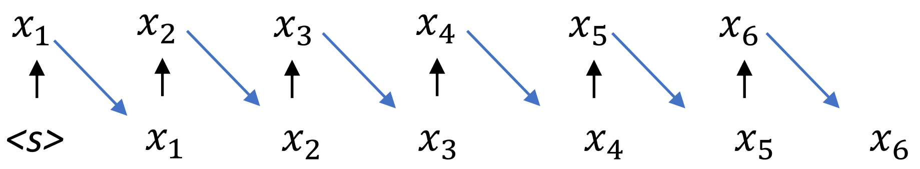


Pre-training Tasks: Language Modeling

Causal Language Modeling (CLM)

Predict the next word/token: predict w_i at the index of i-1 (using the hidden state of w_{i-1})

ELMo/GPT(-2/3)

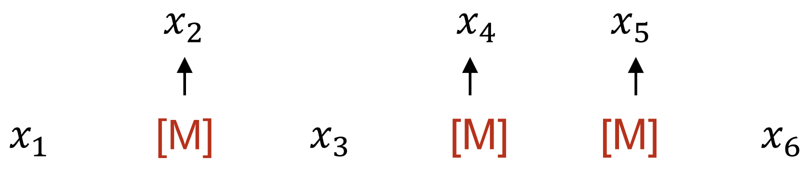


Unidirectional LM

Masked Language Modeling (MLM)

Predict the current word/token: predict w_i at the index of i (using the hidden state of *the placeholder (or masked token) of* w_i)

BERT/RoBERTa/XLNet/UniLM …



Bidirectional LM

The scope of the prediction over context tokens and among context tokens

Pre-training Tasks: Masked Language Modeling

Dependency among masked tokens during prediction

Autoencoding (AE)

BERT/RoBERTa

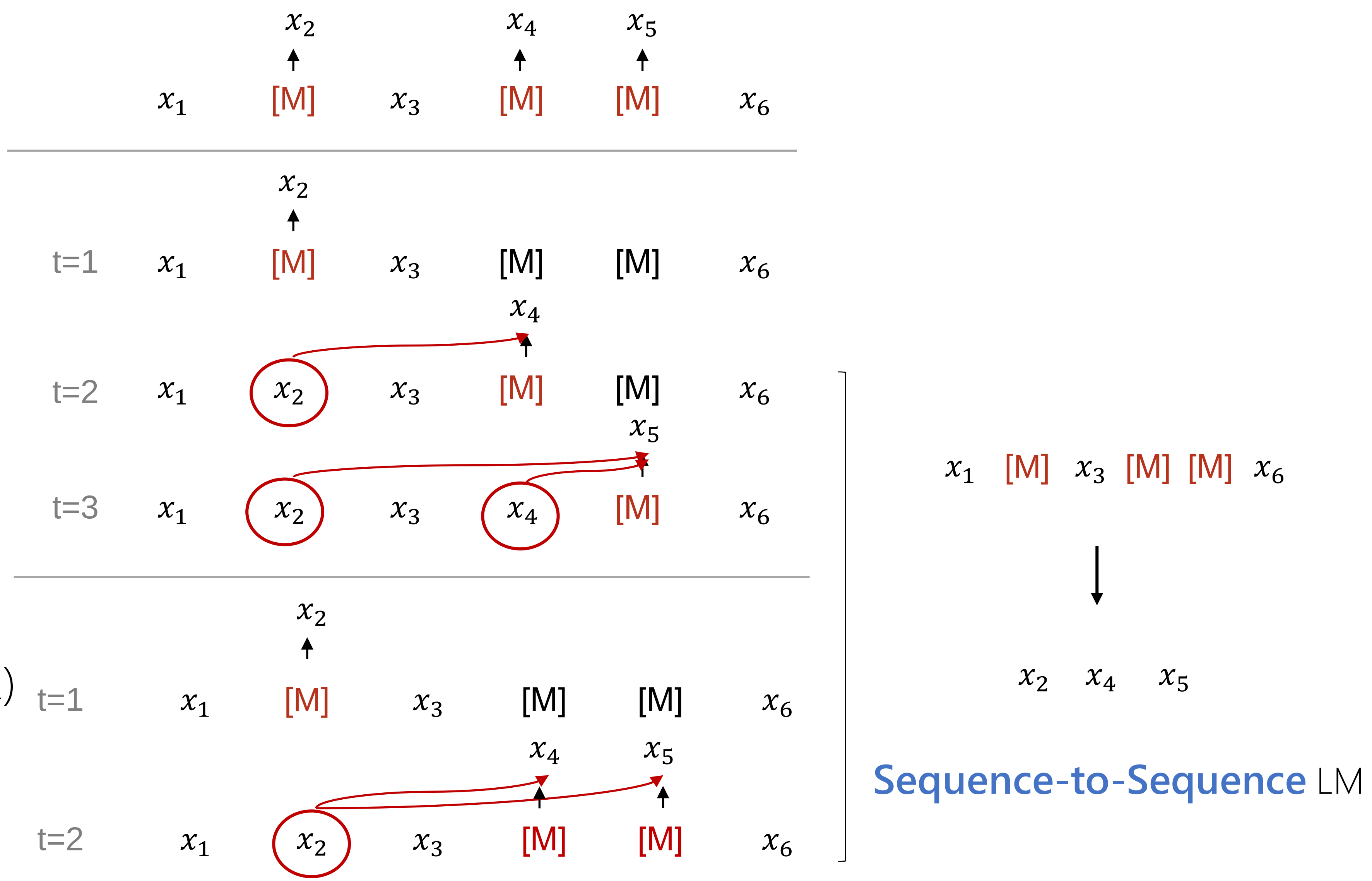
Autoregressive (AR)

XLNet

Partially autoregressive (PAR)

(Part of)

UniLMv2



Pre-training Tasks: Masked Language Modeling

Denoising Autoencoder (DAE)

T5

x_1 [M] x_3 [M] x_6

↓

x_2 [M] x_4 x_5 [M]

BART

x_1 [M] x_3 [M] [M] x_6

↓

x_1 x_2 x_3 x_4 x_5 x_6

Sequence-to-Sequence LM

NLP Tasks & Pre-training Tasks

MLM: Masked Language Modeling
CLM: Causal Language Modeling
DAE: Denoising Autoencoder

NLU

{ NL => X (≠ NL) }

Text (Language E) => { Class, Sequence of class, Scores, …}

Bidirectional

Bidirectional LM

AE
AR
PAR

NLG

{ X (≠ audio/image/video) => NL }

NLG (LM)

Text (Language E) => Text (Language E)

Unidirectional

Unidirectional LM

CLM

NLG (S2S)

{ } => Text (Language E)

NLG (XL)/MT

Text (Language E) => Text (Language F)

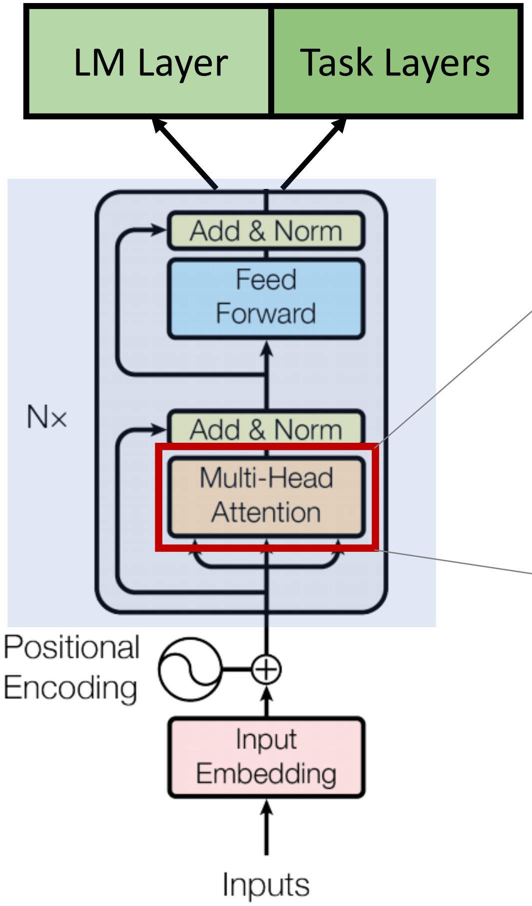
Bidirectional + Unidirectional

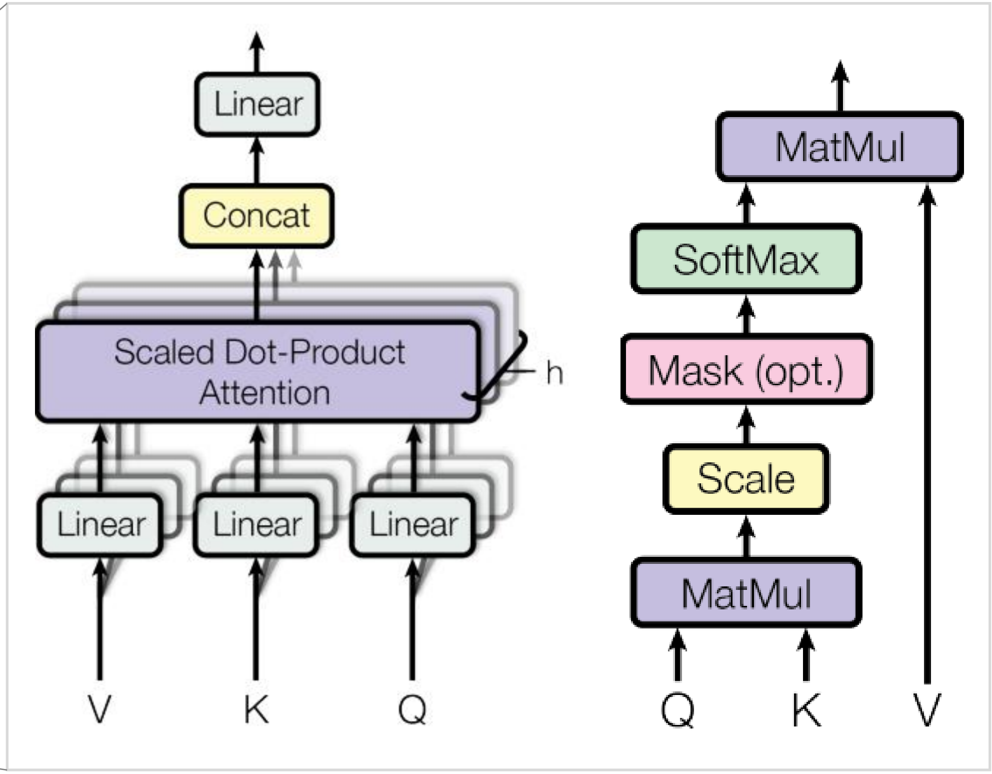
Sequence-to-Sequence LM

DAE
AR/PAR

<u>Context Modeling</u>

Backbone Networks: Multilayer Transformers





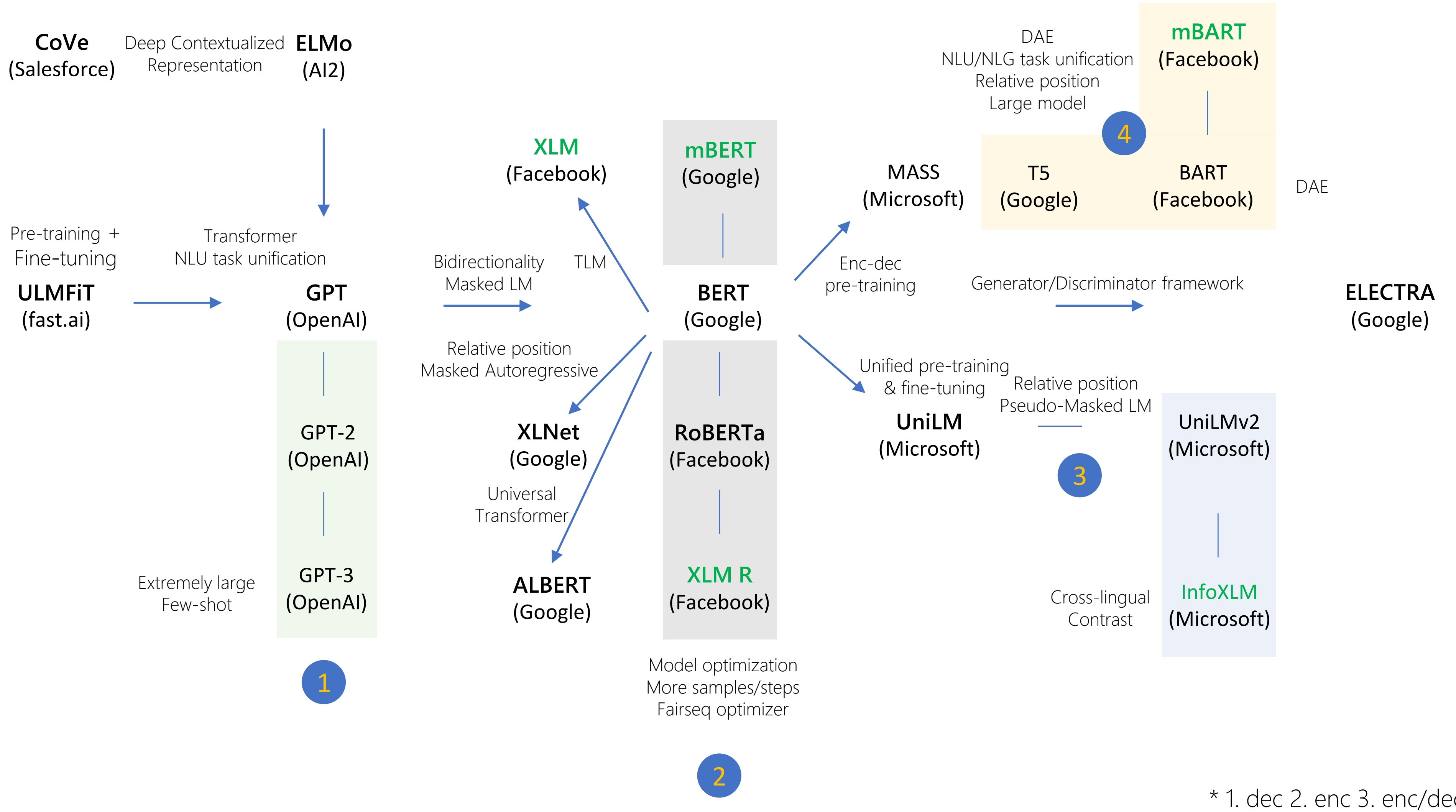
Multi-Head Attention

Scaled Dot-Product Attention

$$\mathrm{MultiHead}(Q,K,V) = \mathrm{Concat}(\mathrm{head}_1, ..., \mathrm{head}_h)W^O$$
$$\text{where } \mathrm{head}_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The Anatomy of Language Model Pre-training



Architectures
- enc/unified, dec, enc-dec

Backbone Networks
- Transformer

Pre-training tasks
- MLM

* 1. dec 2. enc 3. enc/dec 4. enc-dec

UniLM: Unified Language Model Pre-training

05/19

UniLMv1
(NeurIPS'19)

02/20

UniLMv2
(ICML'20)

MiniLM
(NeurIPS'20)

LayoutLM
(KDD'20)

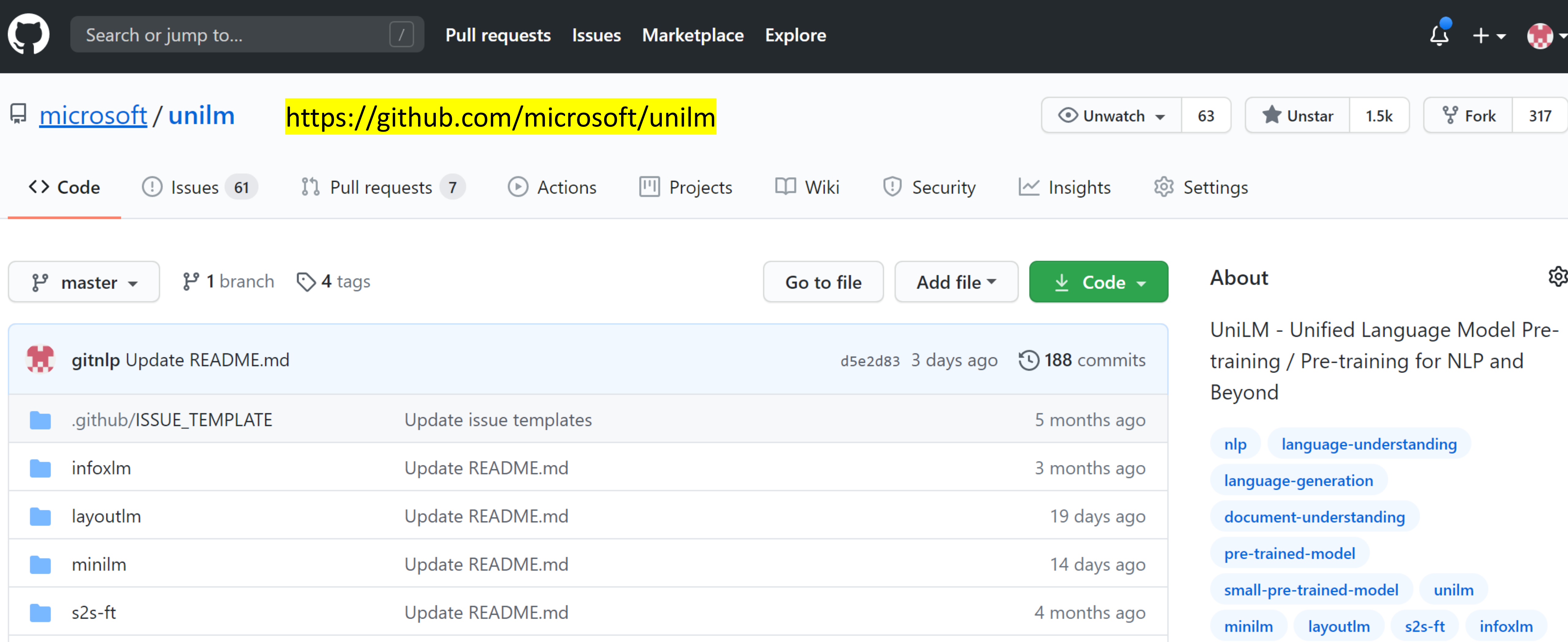
06/20

InfoXLM

(Multilingual/Cross-lingual UniLM)

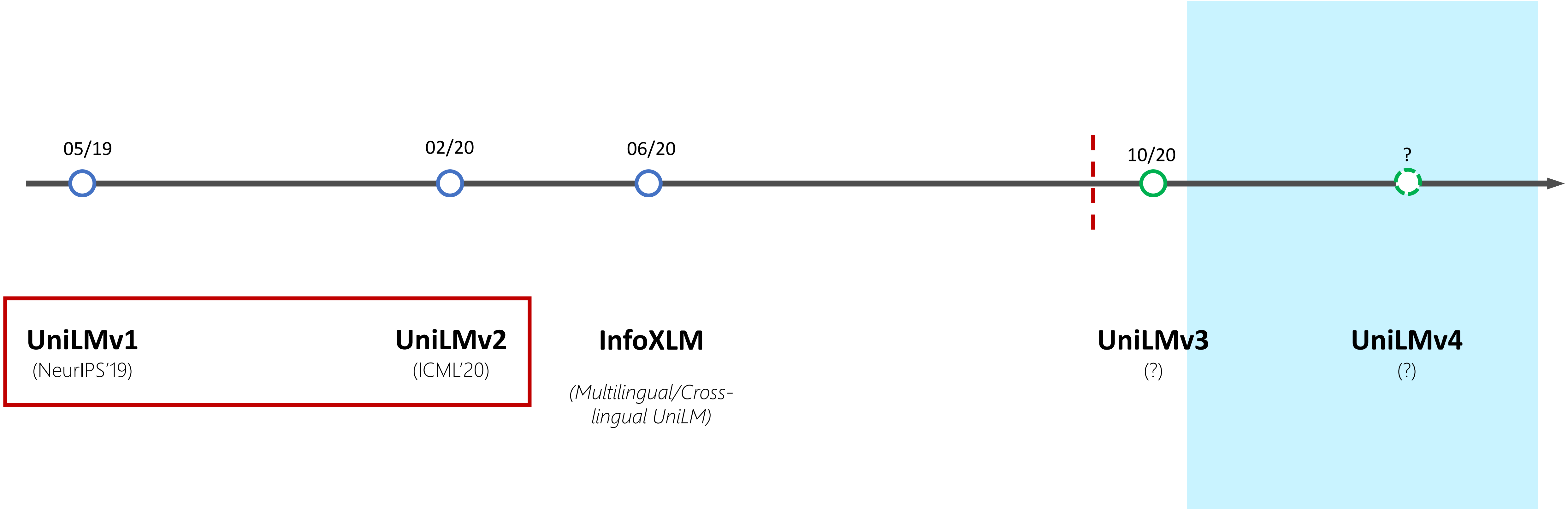
10/20

UniLMv3
(?)



- UniLMv2: Pseudo-Masked Language Models for Unified Language Model Pre-Training. **ICML** 2020.
- Unified Language Model Pre-training for Natural Language Understanding and Generation. **NeurIPS** 2019. // UniLMv1
- MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. **NeurIPS** 2020.
- InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training. Preprint 2020.
- LayoutLM: Pre-training of Text and Layout for Document Image Understanding. **KDD** 2020.

UniLM: Unified Language Model Pre-training



* Unified Language Model Pre-training for Natural Language Understanding and Generation. **NeurIPS** 2019.
* UniLMv2: Pseudo-Masked Language Models for Unified Language Model Pre-Training. **ICML** 2020.

Unified Language Model Pre-training

MLM: Masked Language Modeling
CLM: Causal Language Modeling
DAE: Denoising Autoencoder

NLP Tasks

NLU { NL => X (≠ NL) }

Text (Language E) => { Class, Sequence of class, Score, …}

NLG { X (≠ audio/image/video) => NL }

NLG (LM)

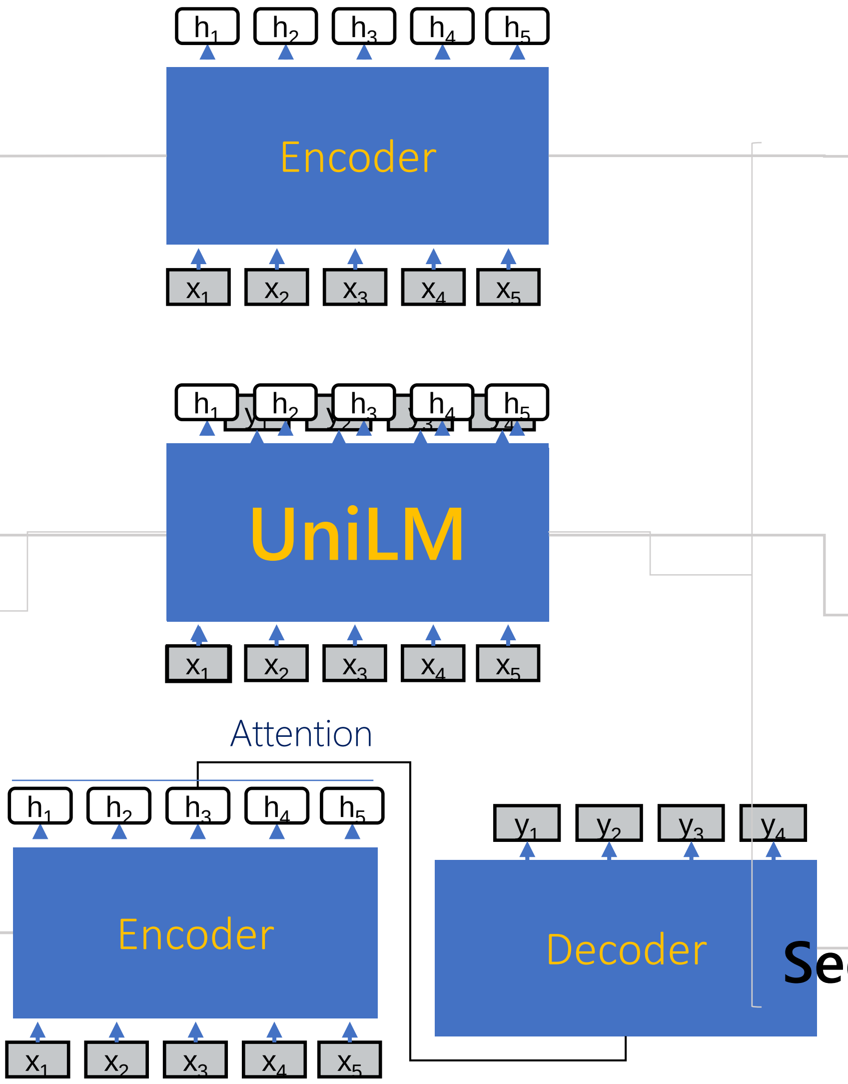
{ } => Text (Language E)

NLG (S2S)

Text (Language E) => Text (Language E)

NLG (XL)/MT

Text (Language E) => Text (Language F)



Pre-training Tasks

Bidirectional LM

AE

ɔBERTa

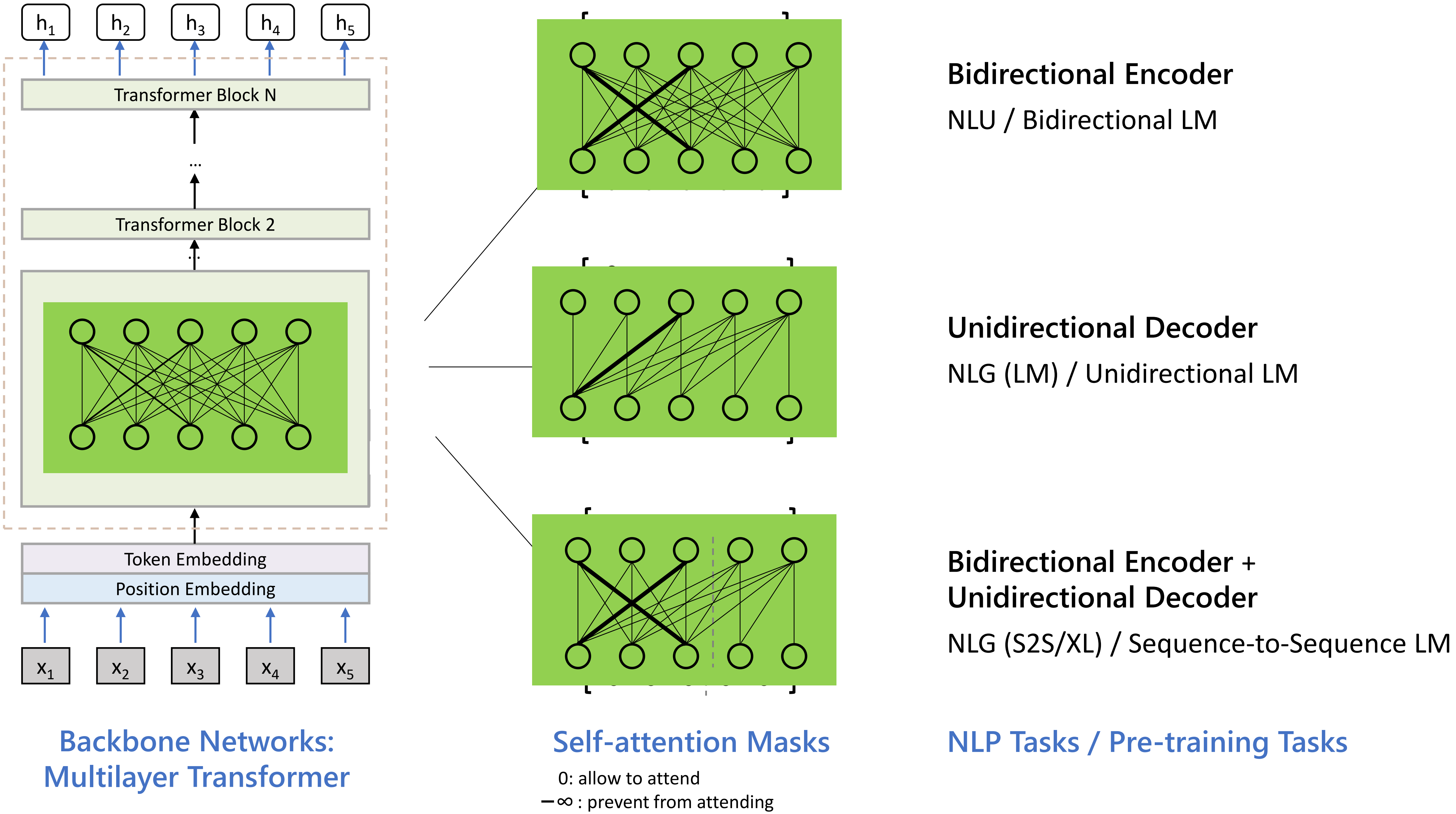
Unidirectional LM

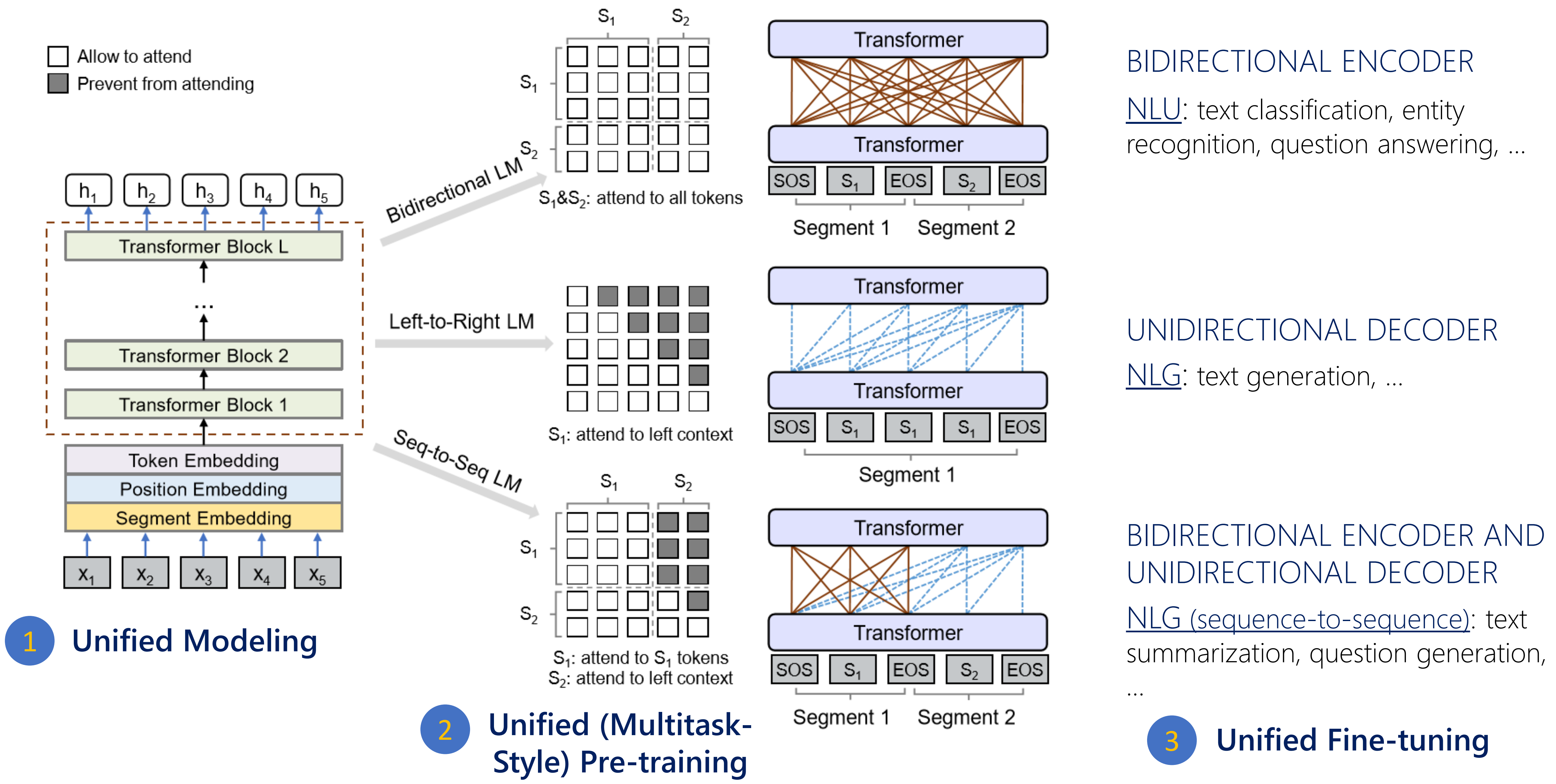
CLM

1/2/3

Sequence-to-Sequence LM

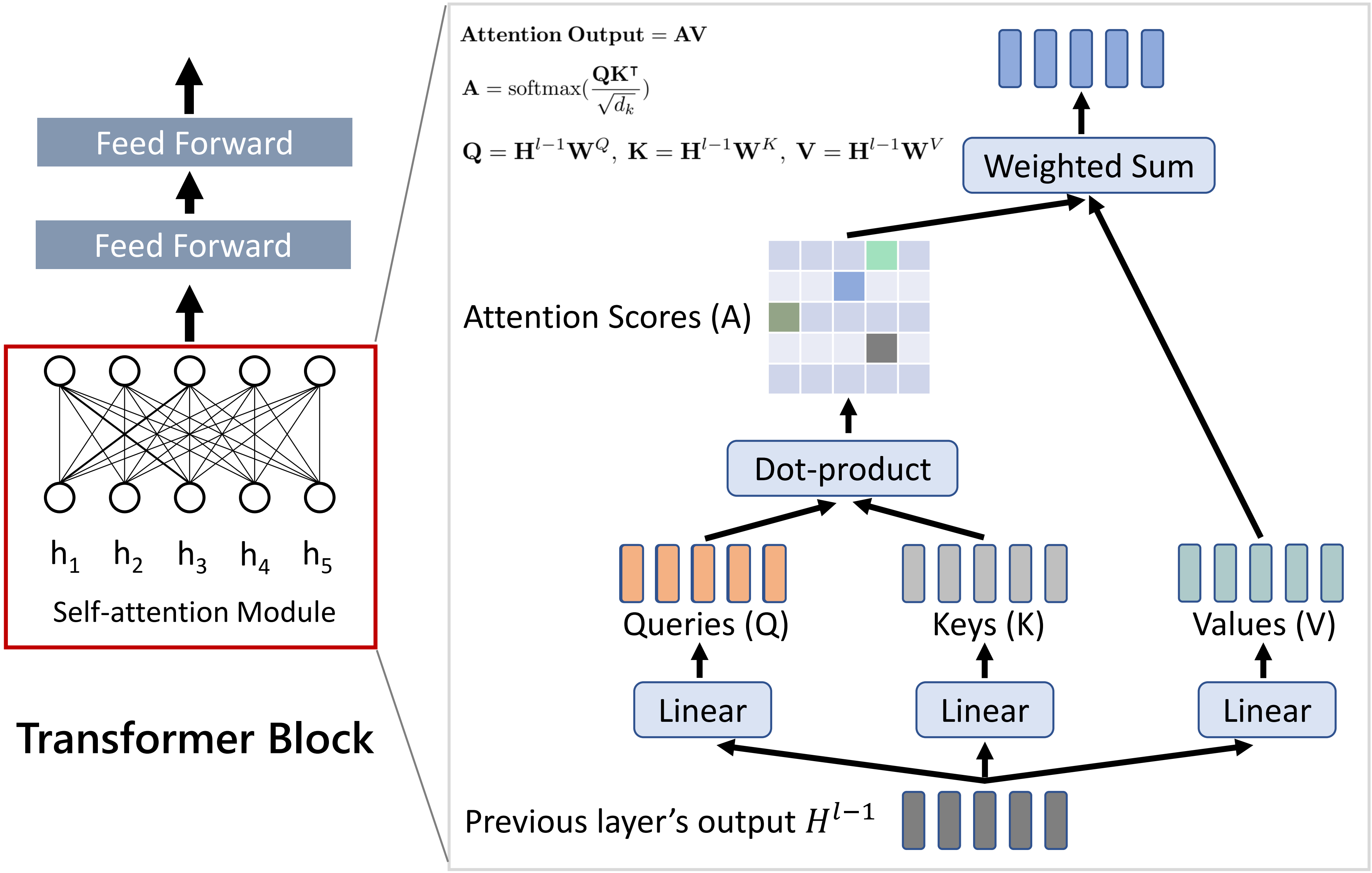
PAR


h1
h2
h3
h4
h5
Transformer Block N
...
Transformer Block 2
Token Embedding
Position Embedding
x1
x2
x3
x4
x5
Backbone Networks:
Multilayer Transformer
Bidirectional Encoder
NLU / Bidirectional LM
Unidirectional Decoder
NLG (LM) / Unidirectional LM
Bidirectional Encoder +
Unidirectional Decoder
NLG (S2S/XL) / Sequence-to-Sequence LM
Self-attention Masks
0: allow to attend
−∞ : prevent from attending
NLP Tasks / Pre-training Tasks



Overview of the Unified Pre-training Framework in UniLM. We use different self-attention masks to control the access to context for each word token in different variants of LMs in pre-training and different NLU/NLG tasks in fine-tuning.

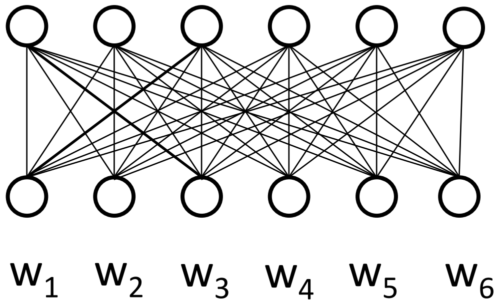
Transformer Block & Self-Attention

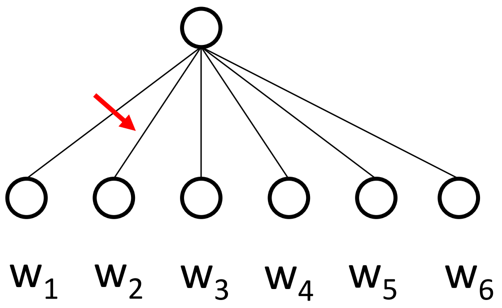


Self-Attention Masks

$$\mathbf{Q} = \mathbf{H}_{L-1}\mathbf{W}_L^Q \qquad \mathbf{K} = \mathbf{H}_{L-1}\mathbf{W}_L^K \qquad \mathbf{V} = \mathbf{H}_{L-1}\mathbf{V}_L$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$





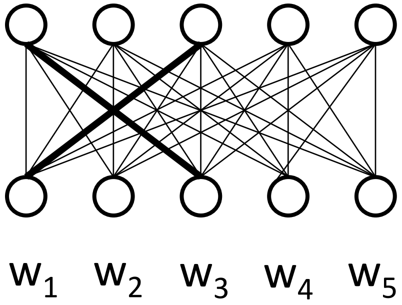
$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{QK}^\intercal}{\sqrt{d_k}} \boxed{+\,\mathbf{M}}\right)V$$

$$\boxed{\mathbf{M}_{ij} = \begin{cases} 0, & \text{allow to attend} \\ -\infty, & \text{prevent from attending} \end{cases}}$$

$$M = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\infty & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Self-Attention Masks for UniLM

Bidirectional LM



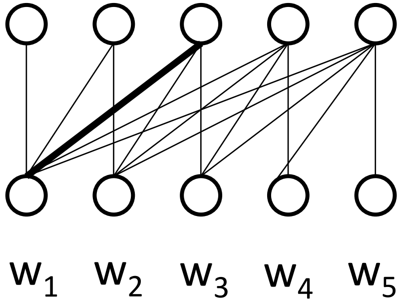
$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Bidirectional Modeling

NLU (Natural Language Understanding)

Unidirectional LM

Left-to-right / Right-to-left

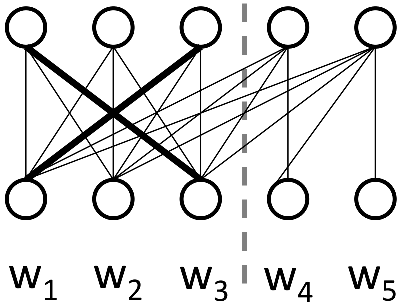


$$\begin{bmatrix} 0 & -\infty & -\infty & -\infty & -\infty \\ 0 & 0 & -\infty & -\infty & -\infty \\ 0 & 0 & 0 & -\infty & -\infty \\ 0 & 0 & 0 & 0 & -\infty \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Unidirectional Modeling

NLG (Unconditional Language Generation)

Sequence-to-Sequence LM



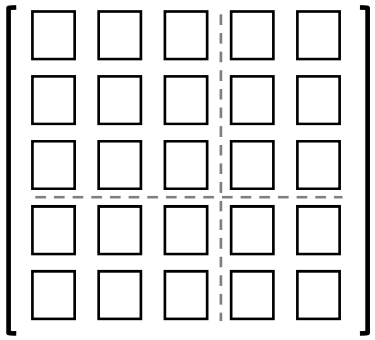
$$\left[\begin{array}{ccc|cc} 0 & 0 & 0 & -\infty & -\infty \\ 0 & 0 & 0 & -\infty & -\infty \\ 0 & 0 & 0 & -\infty & -\infty \\ \hline 0 & 0 & 0 & 0 & -\infty \\ 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

Bidirectional + Unidirectional Modeling

NLG (Sequence-to-Sequence Generation)

The same unified pre-trained model

$+$

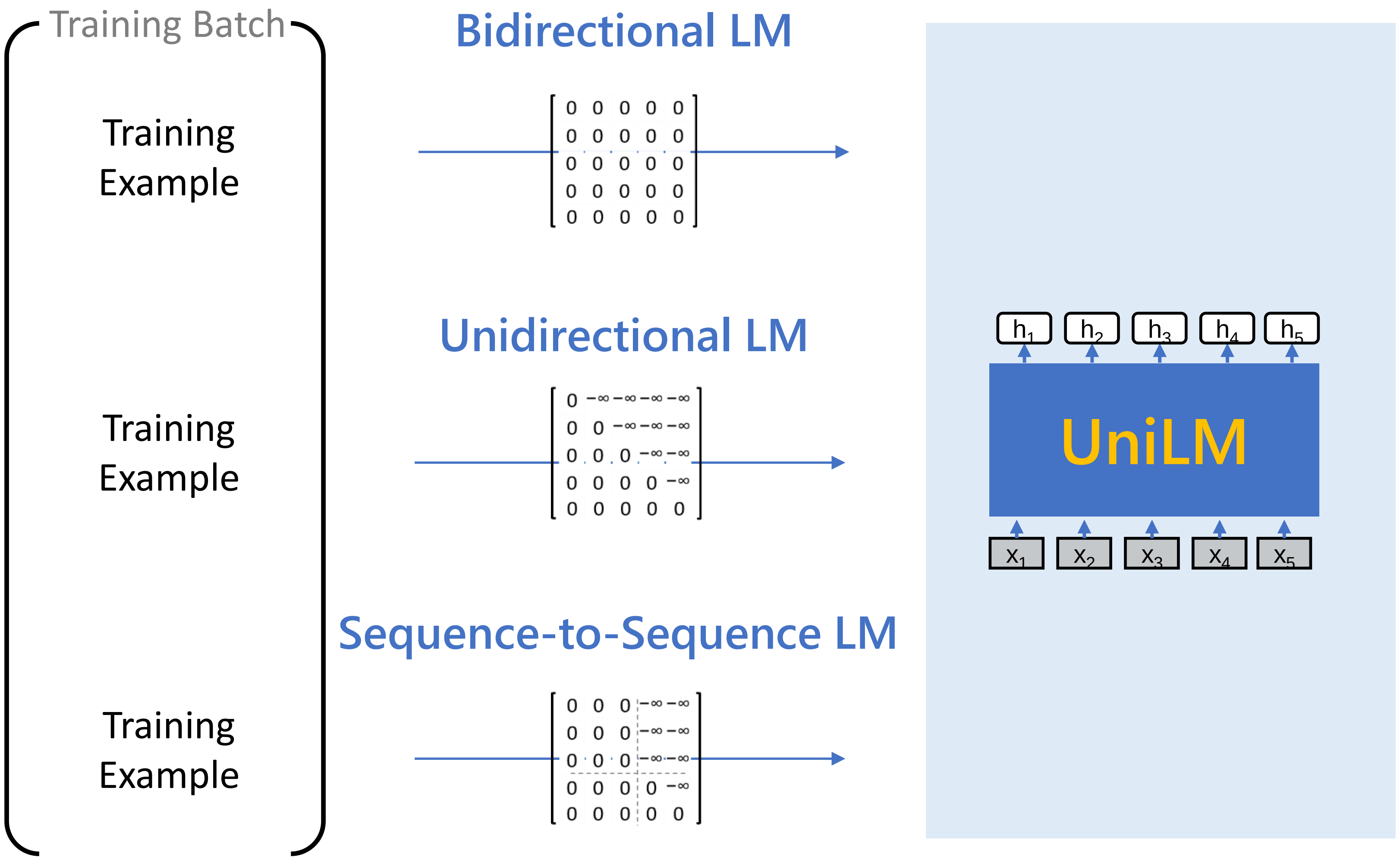


Motivation of UniLMv2

(v1) One training example for each type of LM

- Three types of LMs
- Three forward passes with different self-attention masks

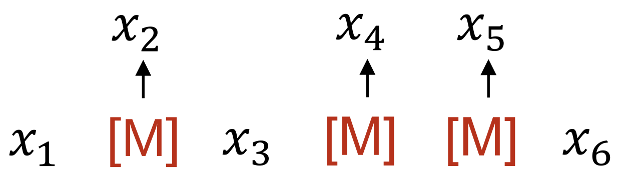
How to train multiple LMs in one forward pass?



Pseudo-Masked Language Model

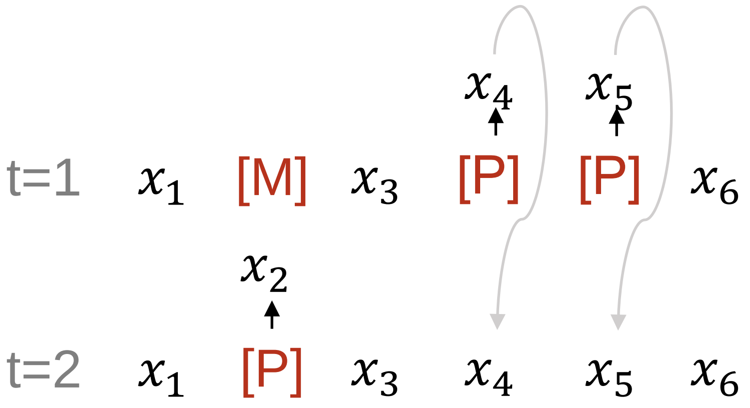
Bidirectional LM Task

1. Bidirectionally encode context tokens
2. Predict the masked spans at the same time



Sequence-to-Sequence LM Task

1. Bidirectionally encode context tokens
2. Predict the masked spans one by one
    1. Predict x_4, x_5
    2. Encode x_4, x_5 (i.e., fill in what we have predicted)
    3. Predict x_2



Pseudo-Masked Language Model

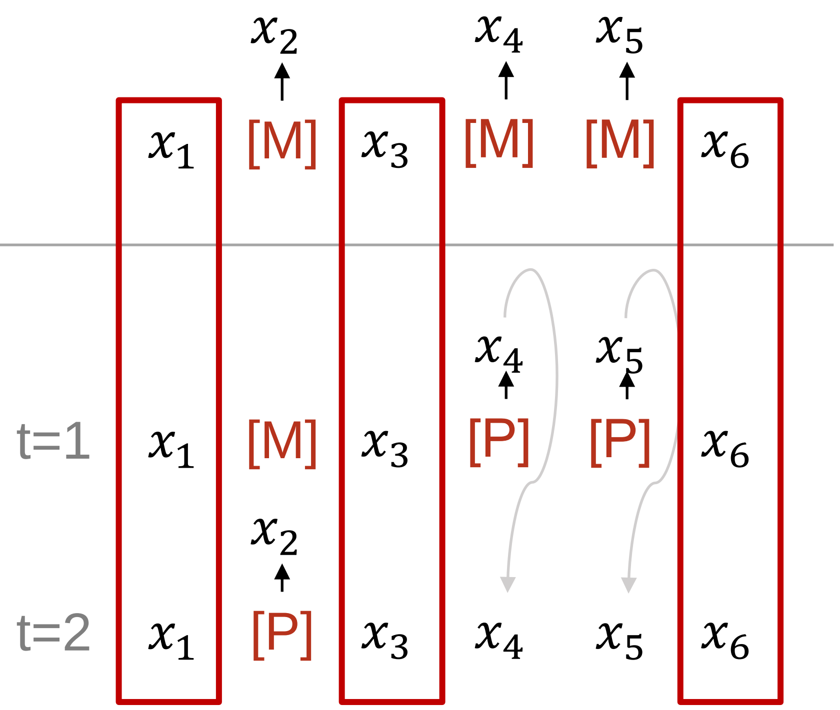
Observation #1: context encoding can be reused

Bidirectional LM Task

1. Bidirectionally encode context tokens
2. Predict the masked spans at the same time

Sequence-to-Sequence LM Task

1. Bidirectionally encode context tokens
2. Predict the masked spans one by one
    1. Predict x_4, x_5
    2. Encode x_4, x_5 (i.e., fill in what we have predicted)
    3. Predict x_2

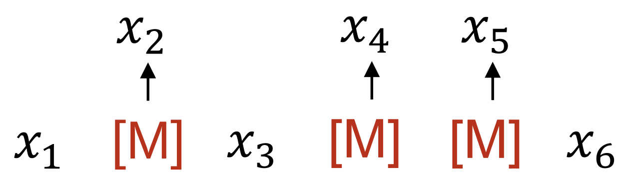


Pseudo-Masked Language Model

Observation #1: context encoding can be reused
Observation #2: masked positions have three roles

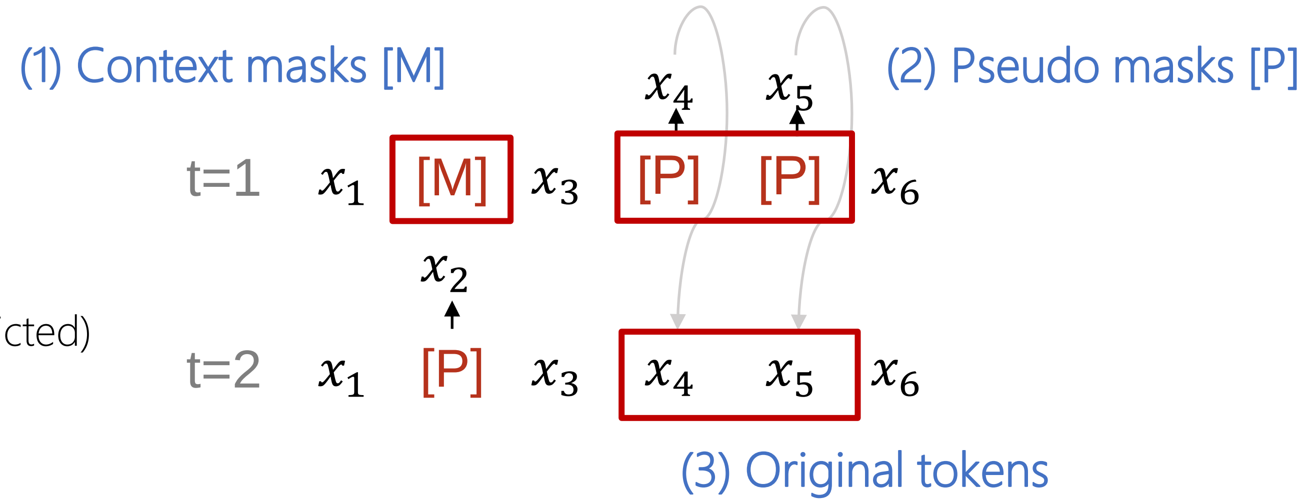
Bidirectional LM Task

1. Bidirectionally encode context tokens
2. Predict the masked spans at the same time



Sequence-to-Sequence LM Task

1. Bidirectionally encode context tokens
2. Predict the masked spans one by one
    1. Predict x_4, x_5
    2. Encode x_4, x_5 (i.e., fill in what we have predicted)
    3. Predict x_2

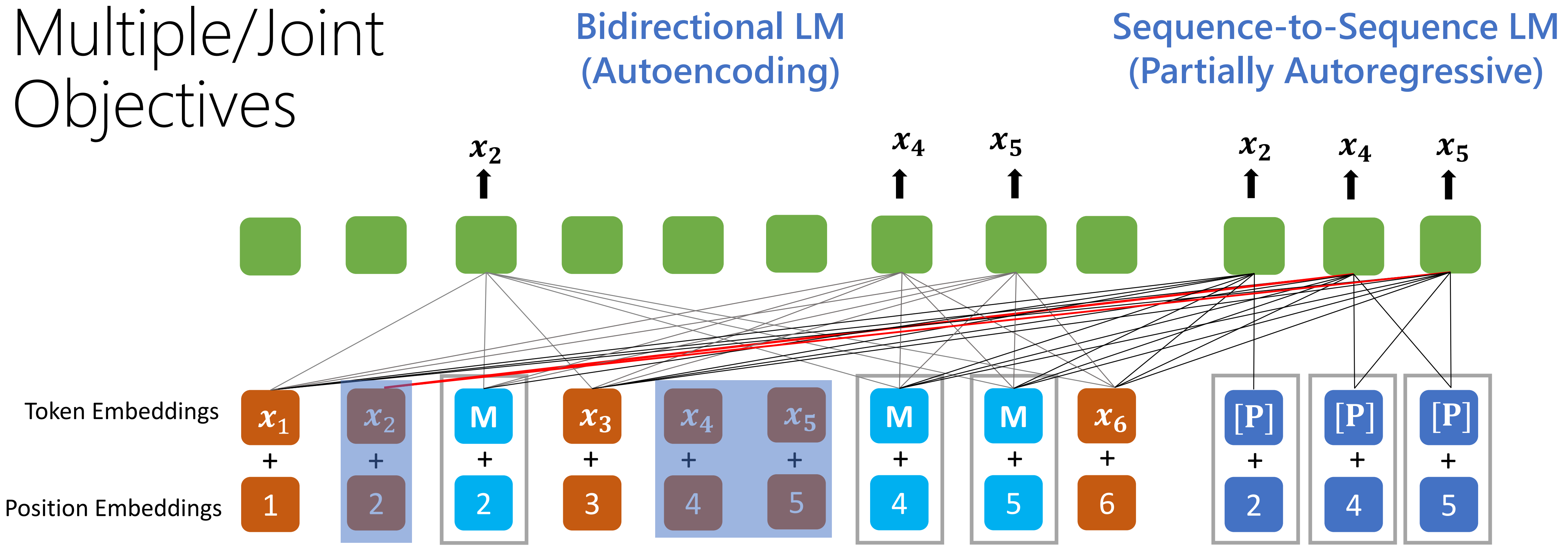


Pseudo-Masked Language Model

- A general framework to
  - **efficiently** realizes different pre-training objectives
    - AE (autoencoding): BERT/RoBERTa
    - AR (autoregressive): XLNet
    - PAR (partially autoregressive): New in UniLMv2
    - AE + AR: BERT + XLNet
    - **AE + PAR: UniLMv2 (unified pre-training), performs the best**
  - **effectively** learn different word dependencies
    - Between context and mask predictions
    - Between mask predictions

in **one forward pass**

Multiple/Joint Objectives

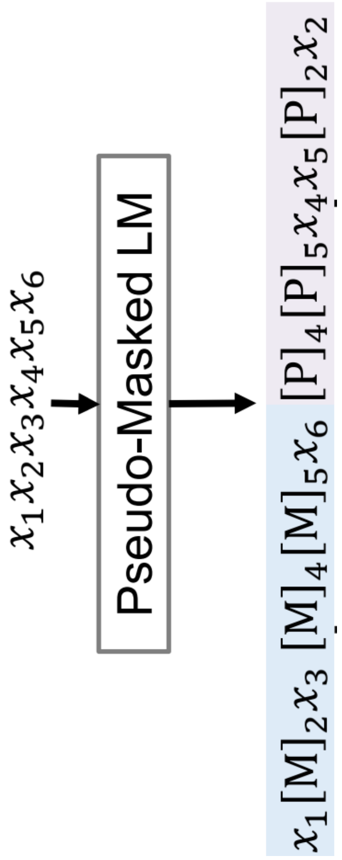


UniLM (v2): **Unified pre-training** of bi-directional LM (via autoencoding) and sequence-to-sequence LM (via **partially autoregressive**) with **Pseudo-Masked Language Model** for **language understanding and generation**

- Transformer/Self-attention treats tokens with the same position embeddings as the *same "token"* at that position
- Pseudo-masked LM can be used to efficiently realize different pre-training objectives, such as AE (autoencoding), AR (autoregressive), PAR (partially autoregressive) , AE + AR, and AE + PAR, among which AE + PAR performs the best

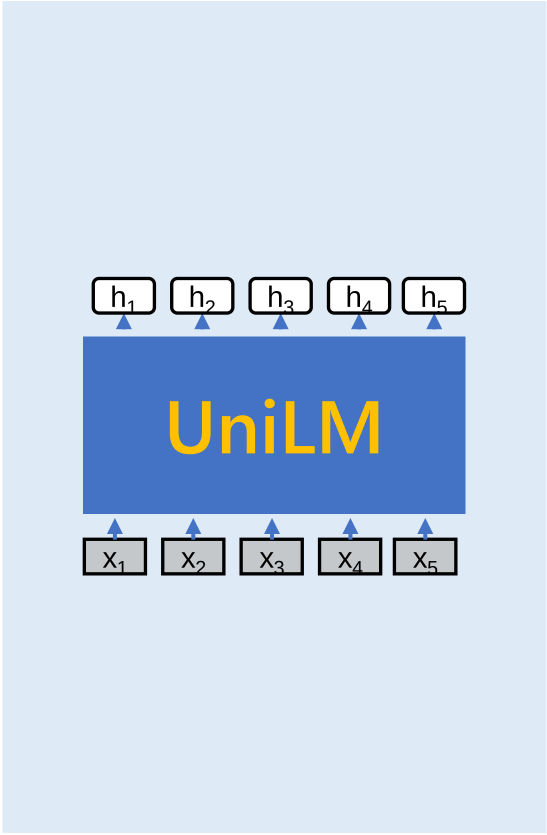
Self-Attention Masks for UniLMv2

Unified pre-training of bi-directional LM (via autoencoding) and sequence-to-sequence LM (via partially autoregressive) with **one training sample (forward pass)**



+

$$
\begin{array}{cc}
 & \begin{array}{cccccccccccc} 1 & 2 & 2 & 2 & 3 & 4 & 5 & 4 & 5 & 4 & 5 & 6 \\ x_1 & x_2 & [P] & [M] & x_3 & x_4 & x_5 & [P] & [P] & [M] & [M] & x_6 \end{array} \\
\begin{array}{cc} 1 & x_1 \\ 2 & x_2 \\ 2 & [P] \\ 2 & [M] \\ 3 & 3 \\ 4 & 4 \\ 5 & 5 \\ 4 & [P] \\ 5 & [P] \\ 4 & [M] \\ 5 & [M] \\ 6 & x_6 \end{array} &
\left[\begin{array}{cccccccccccc}
0 & -\infty & -\infty & 0 & 0 & -\infty & -\infty & -\infty & -\infty & 0 & 0 & 0 \\
0 & 0 & -\infty & 0 & 0 & -\infty & -\infty & -\infty & -\infty & 0 & 0 & 0 \\
0 & -\infty & 0 & 0 & 0 & -\infty & -\infty & -\infty & -\infty & 0 & 0 & 0 \\
0 & -\infty & -\infty & 0 & 0 & -\infty & -\infty & -\infty & -\infty & 0 & 0 & 0 \\
0 & -\infty & -\infty & 0 & 0 & -\infty & -\infty & -\infty & -\infty & 0 & 0 & 0 \\
0 & -\infty & -\infty & 0 & 0 & 0 & 0 & -\infty & -\infty & 0 & 0 & 0 \\
0 & -\infty & -\infty & 0 & 0 & 0 & 0 & -\infty & -\infty & 0 & 0 & 0 \\
0 & 0 & -\infty & 0 & 0 & -\infty & -\infty & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -\infty & 0 & 0 & -\infty & -\infty & 0 & 0 & 0 & 0 & 0 \\
0 & -\infty & -\infty & 0 & 0 & -\infty & -\infty & -\infty & -\infty & 0 & 0 & 0 \\
0 & -\infty & -\infty & 0 & 0 & -\infty & -\infty & -\infty & -\infty & 0 & 0 & 0 \\
0 & -\infty & -\infty & 0 & 0 & -\infty & -\infty & -\infty & -\infty & 0 & 0 & 0
\end{array}\right]
\end{array}
$$



UniLMv2 Base for NLU Tasks

Model	SQuAD v1.1		SQuAD v2.0	
	F1	EM	F1	EM
BERT	88.5	80.8	76.3	73.7
XLNet	-	-	-	80.2
RoBERTa	91.5	84.6	83.7	80.5
UniLMv2	**93.1**	**87.1**	**86.1**	**83.3**

Results of **BASE-size** pre-trained models on the **SQuAD v1.1/v2.0** development sets. We report F1 scores and exact match (EM) scores. Results of UniLMv2 are averaged over five runs.

Model	MNLI	SST-2	MRPC	RTE	QNLI	QQP	STS	CoLA
	Acc	Acc	Acc	Acc	Acc	Acc	PCC	MCC
BERT	84.5	93.2	87.3	68.6	91.7	91.3	89.5	58.9
XLNet	86.8	94.7	88.2	74.0	91.7	91.4	89.5	60.2
RoBERTa	87.6	94.8	90.2	78.7	92.8	**91.9**	**91.2**	63.6
UniLMv2	**88.5**	**95.1**	**91.8**	**81.3**	**93.5**	91.7	91.0	**65.2**

Results of **BASE-size** models on the development set of the **GLUE benchmark**. We report Matthews correlation coefficient (MCC) for CoLA, Pearson correlation coefficient (PCC) for STS, and accuracy (Acc) for the rest. Metrics of UniLMv2 are averaged over five runs for the tasks.

UniLMv2 Base/Large for NLG Tasks

Outperforming T5(11B), BART, Pegasus, w/ 3% of T5 parameters, 20% of T5 (4% of Pegasus) data

Model	#Param	Corpus	CNN/DailyMail RG-1/RG-2/RG-L	XSum RG-1/RG-2/RG-L	Gigaword RG-1/RG-2/RG-L
Without pre-training					
PtrNet (See et al., 2017)	-	-	39.53/17.28/36.38	28.10/8.02/21.72	-
Fine-tuning BASE-*size pre-trained models*					
$\text{MASS}_{\text{BASE}}$ (Song et al., 2019)	123M	-	42.12/19.50/39.01	39.75/17.24/31.95	38.73/19.71/35.96
BERTSumAbs (Liu & Lapata, 2019)	156M	16GB	41.72/19.39/38.76	38.76/16.33/31.15	-
$\text{ERNIE-GEN}_{\text{BASE}}$ (Xiao et al., 2020)	110M	16GB	42.30/19.92/39.68	-	38.83/20.04/36.20
T5_{BASE} (Raffel et al., 2019)	220M	750GB	42.05/20.34/39.40	-	-
UniLMv2	110M	160GB	**43.89/21.05/41.02**	**44.67/21.78/36.81**	-
Fine-tuning LARGE-*size pre-trained models*					
$\text{UniLMv1}_{\text{LARGE}}$ (Dong et al., 2019)	340M	16GB	43.08/20.43/40.34	-	38.90/20.05/36.00
$\text{ERNIE-GEN}_{\text{LARGE}}$ (Xiao et al., 2020)	340M	16GB	44.02/21.17/41.26	-	39.25/20.25/36.53
$\text{BART}_{\text{LARGE}}$ (Lewis et al., 2019)	400M	160GB	44.16/21.28/40.90	45.14/22.27/37.25	-
ProphetNet (Yan et al., 2020)	400M	160GB	44.20/21.17/41.30	-	39.51/20.42/36.69
$\text{PEGASUS}_{\text{C4}}$ (Zhang et al., 2019)	568M	750GB	43.90/21.20/40.76	45.20/22.06/36.99	38.75/19.96/36.14
$\text{PEGASUS}_{\text{HugeNews}}$ (Zhang et al., 2019)	568M	3800GB	44.17/21.47/41.11	47.21/**24.56**/39.25	39.12/19.86/36.24
T5_{11B} (Raffel et al., 2019)	11B	750GB	43.52/21.55/40.69	-	-
UniLMv2	340M	160GB	**44.79/21.98/41.93**	**47.58**/24.35/**39.50**	**39.73/20.70/36.89**

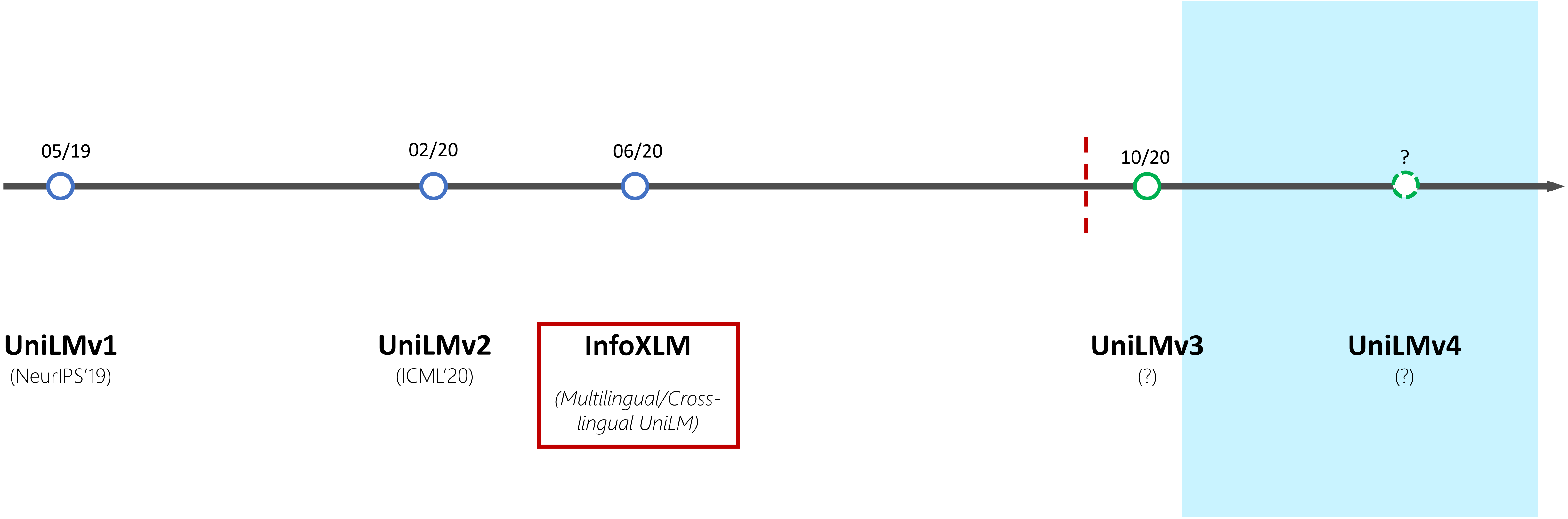
Abstractive summarization results on CNN/DailyMail and XSum. The evaluation metric is the F1 version of ROUGE (RG) scores. We also include the number of parameters (#Param) for the methods using pre-trained models.

UniLMv2 Base/Large for NLG Tasks

Model	#Param	Corpus	Official Split BLEU-4 / MTR / RG-L	Reversed Split BLEU-4 / MTR / RG-L
Without pre-training				
(Du & Cardie, 2018)	-	-	15.16 / 19.12 / -	-
(Zhao et al., 2018)	-	-	-	16.38 / 20.25 / 44.48
(Zhang & Bansal, 2019)	-	-	18.37 / 22.65 / 46.68	20.76 / 24.20 / 48.91
Fine-tuning BASE-*size pre-trained models*				
ERNIE-GEN$_{\text{BASE}}$ (Xiao et al., 2020)	110M	16GB	22.28 / 25.13 / 50.58	23.52 / 25.61 / 51.45
UNILMV2	110M	160GB	**24.70 / 26.33 / 52.13**	**26.30 / 27.09 / 53.19**
Fine-tuning LARGE-*size pre-trained models*				
UNILMV1$_{\text{LARGE}}$ (Dong et al., 2019)	340M	16GB	22.12 / 25.06 / 51.07	23.75 / 25.61 / 52.04
ERNIE-GEN$_{\text{LARGE}}$ (Xiao et al., 2020)	340M	16GB	24.03 / 26.31 / 52.36	25.57 / 26.89 / 53.31
ProphetNet (Yan et al., 2020)	400M	16GB	25.01 / 26.83 / 52.57	26.72 / 27.64 / 53.79
UNILMV2	340M	160GB	**25.97 / 27.33 / 53.43**	**27.12 / 27.95 / 54.25**

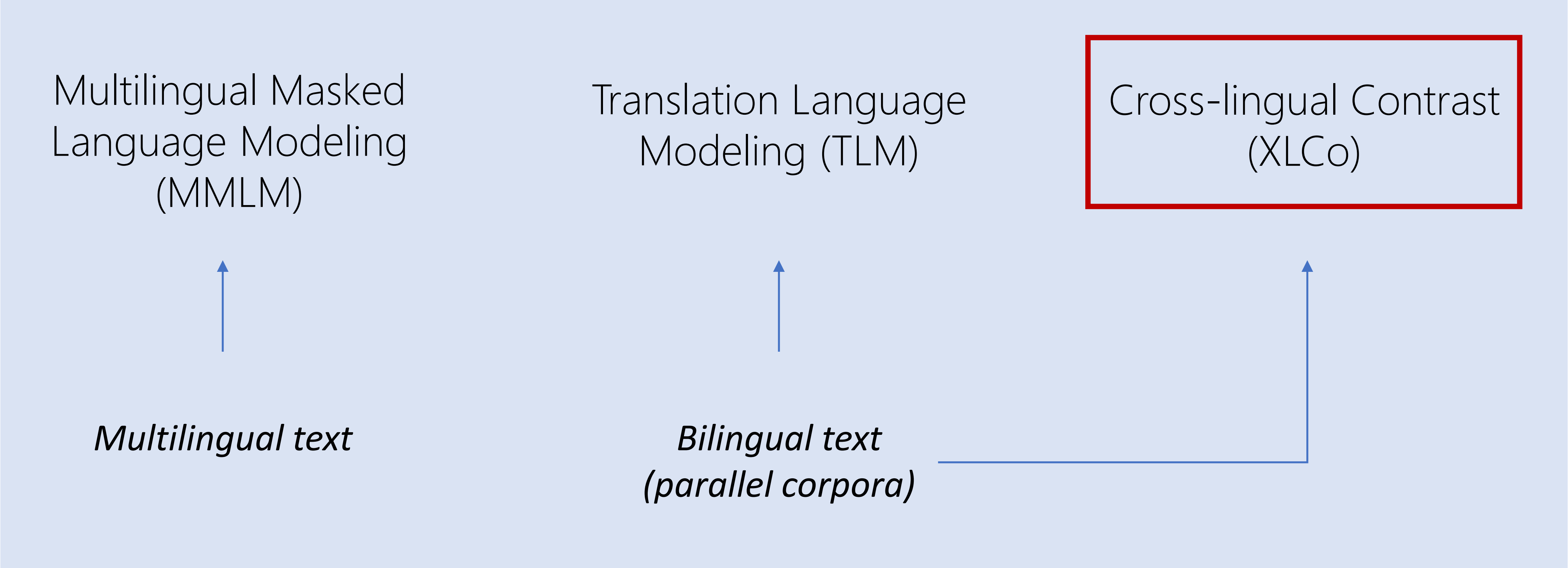
Question generation results on SQuAD v1.1. MTR is short for METEOR, and RG for ROUGE. The official split is from (Du & Cardie, 2018), while the reversed split is the same as in (Zhao et al., 2018).

UniLM: Unified Language Model Pre-training



* InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training. Preprint 2020.

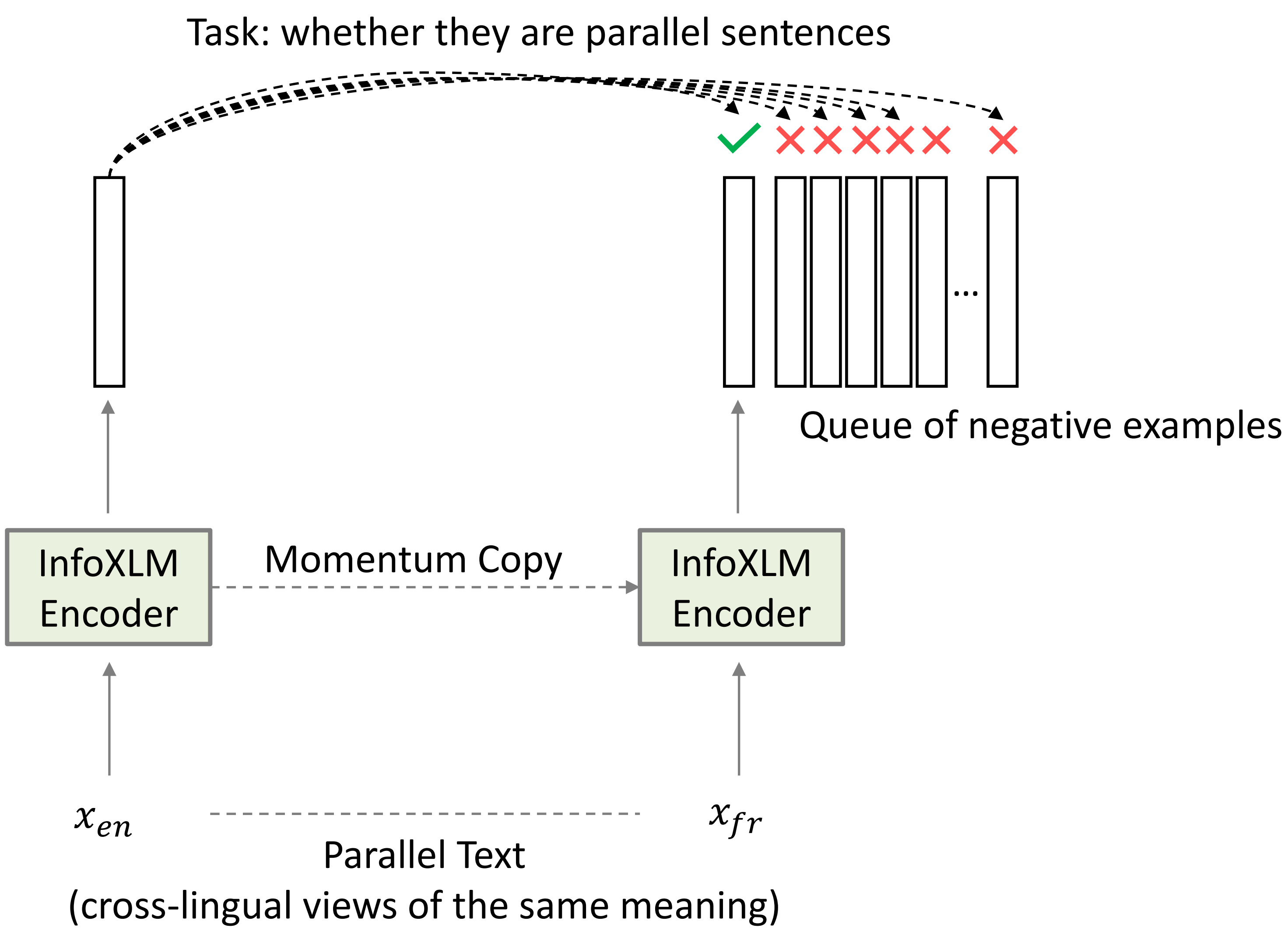
InfoXLM – Cross-lingual Language Model Pre-training



Unified Information Theoretic Framework

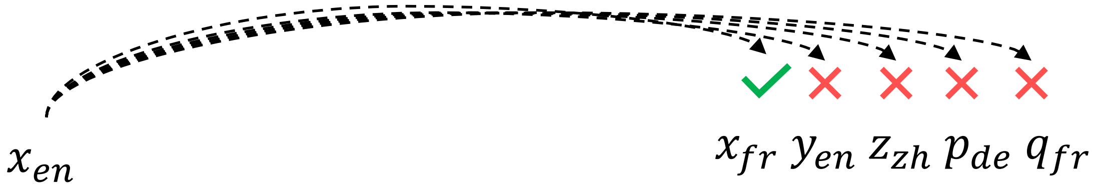
Cross-lingual Contrast (XLCo)

Maximize the InfoNCE (Oord, 2018) lower bound of mutual information $I(x_{en}; x_{fr})$



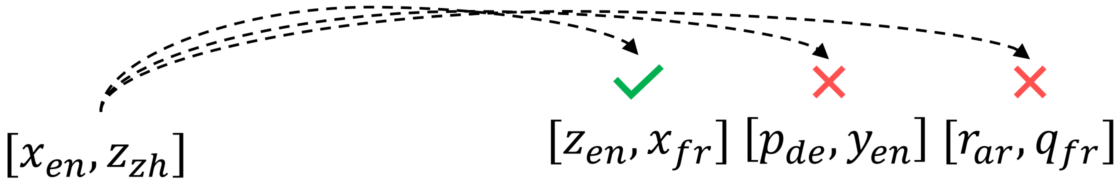
Mixup Contrast

- Because parallel sentences are always in different languages, the model easily knows y_{en} is a negative example of x_{en}
  - Problem: sentences in the same language are "dissimilar" even if they have similar meanings



$(x_., x_.)$ is the correct translation pair

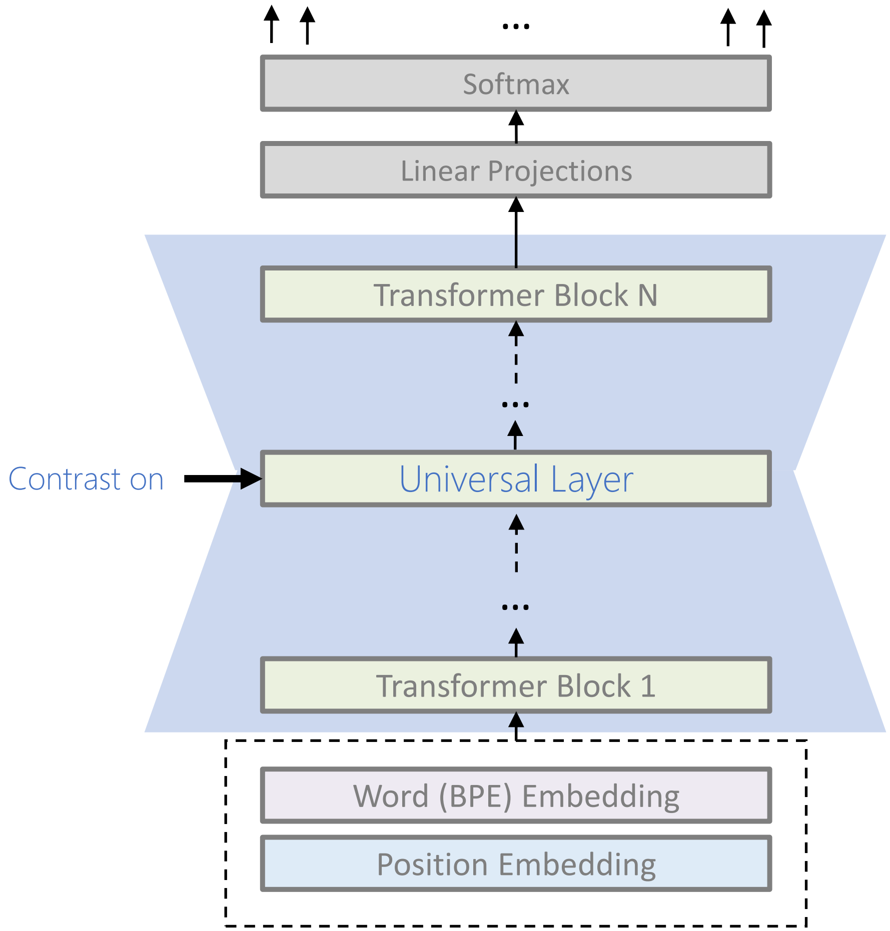
- Solution: mixup contrast
  - Randomly mixup translation pairs and shuffle their orders
  - Encourage multilingual representations blend in together



$(x_., x_.)$ $(z_., z_.)$ are the correct translation pairs

Contrast on Universal Layer

- Multilingual masked language modeling encourages bottleneck structure
  - Language-specific (bottom layers) -> language-agnostic (middle layers) -> language-specific (upper layers)
- Contrastive learning is put on the middle layer to avoid violating the universal space



InfoXLM/T-ULRv2 Achieves #1 on XTREME Benchmark

- Cover four paradigms: classification, structured prediction, question answering, retrieval
- Average over 9 datasets, and 40 languages

Rank	Model	Participant	Affiliation	Attempt Date	Avg	Sentence-pair Classification	Structured Prediction	Question Answering	Sentence Retrieval
0		Human	-	-	93.3	95.1	97.0	87.8	-
1	T-ULRv2 + StableTune	Turing	Microsoft	Oct 7, 2020	80.7	88.8	75.4	72.9	89.3
2	VECO	DAMO NLP Team	Alibaba	Sep 29, 2020	77.2	87.0	70.4	68.0	88.1
3	FILTER	Dynamics 365 AI Research	Microsoft	Sep 8, 2020	77.0	87.5	71.9	68.5	84.4
4	X-STILTs	Phang et al.	New York University	Jun 17, 2020	73.5	83.9	69.4	67.2	76.5
5	XLM-R (large)	XTREME Team	Alphabet, CMU	-	68.2	82.8	69.0	62.3	61.6
6	mBERT	XTREME Team	Alphabet, CMU	-	59.6	73.7	66.3	53.8	47.7
7	MMTE	XTREME Team	Alphabet, CMU	-	59.3	74.3	65.3	52.3	48.9
8	RemBERT	Anonymous2	Anonymous2	-	56.1	84.1	73.3	68.6	-
9	XLM	XTREME Team	Alphabet, CMU	-	55.8	75.0	65.6	43.9	44.7

XTREME Benchmark (CMU & Google): https://sites.research.google/xtreme/

Pre-training for NLP: Revisit

BUILDING SOTA PRE-TRAINED MODELS

Modeling

Network structures, pre-training tasks & objectives

+

Data

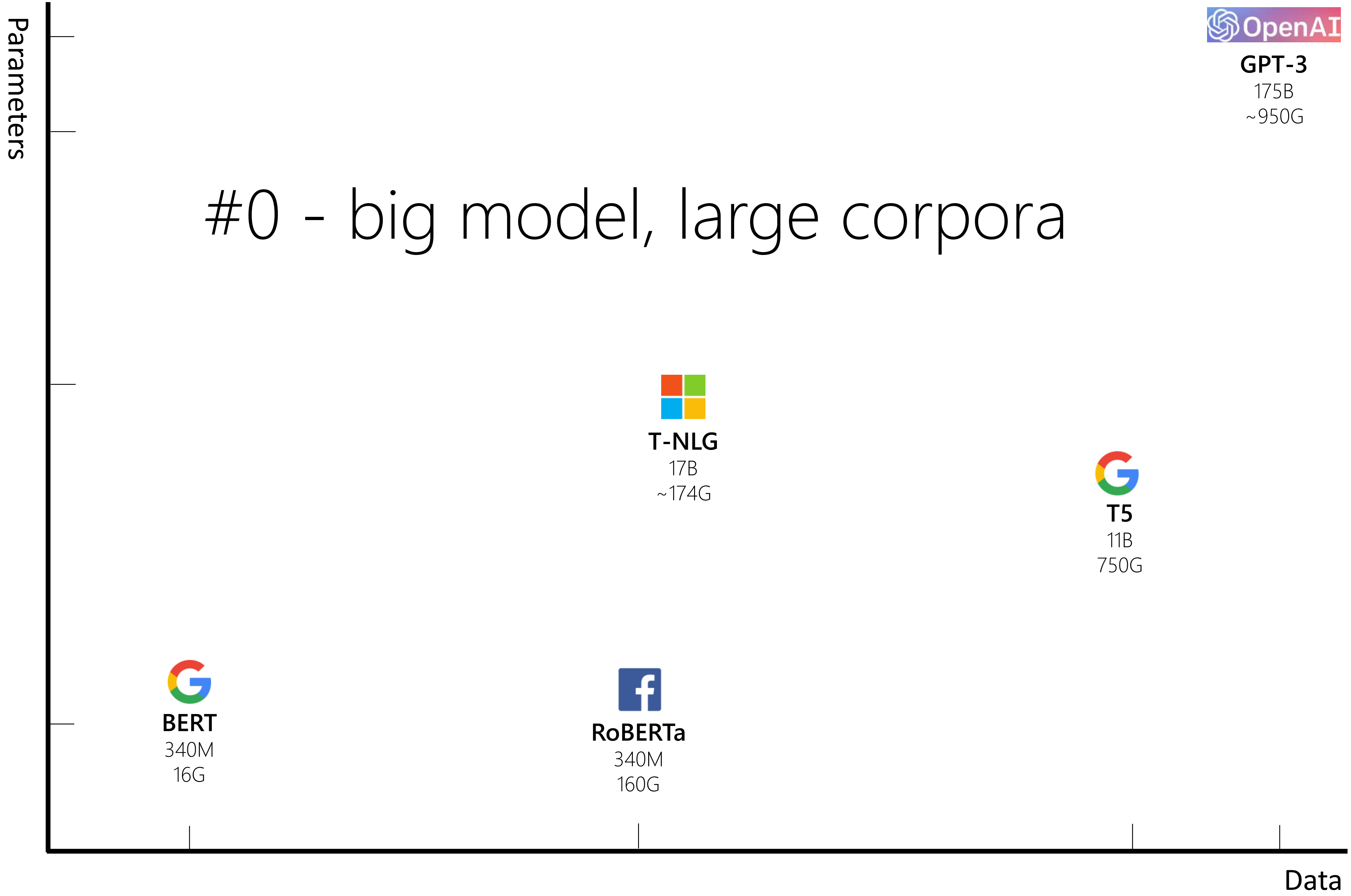
Massive-scale (unlimited) unlabeled corpora

+

System & Infra

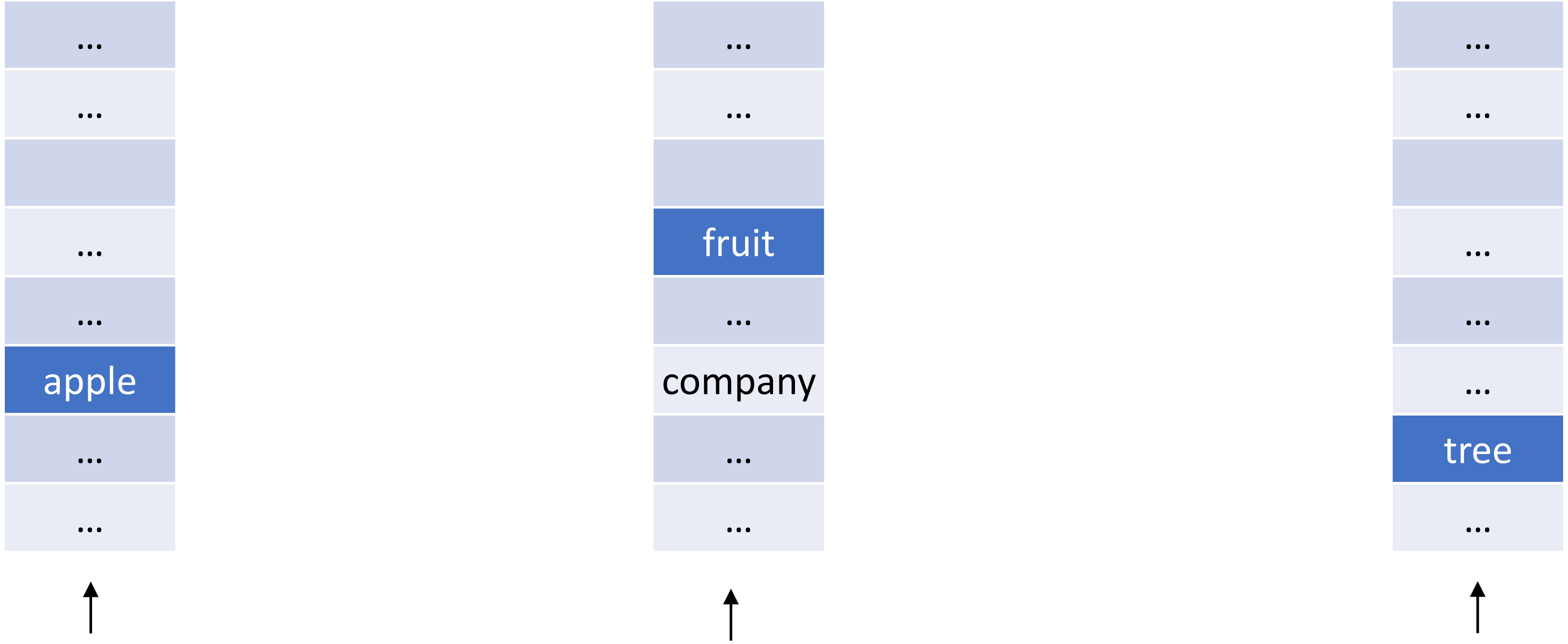
Large-scale training & optimization

#0 - big model, large corpora



#1 - Pre-training tasks: Masked Language Modeling

DAE (enc-dec), TLM (cross-lingual), …



An [MASK] is a sweet, edible [MASK] produced by an apple [MASK]

Masked LM and the Masking Procedure Assuming the unlabeled sentence is `my dog is hairy`, and during the random masking procedure we chose the 4-th token (which corresponding to `hairy`), our masking procedure can be further illustrated by

15% of tokens

- 80% of the time: Replace the word with the `[MASK]` token, e.g., `my dog is hairy` → `my dog is [MASK]`
- 10% of the time: Replace the word with a random word, e.g., `my dog is hairy` → `my dog is apple`
- 10% of the time: Keep the word unchanged, e.g., `my dog is hairy` → `my dog is hairy`. The purpose of this is to bias the representation towards the actual observed word.

* Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2019. *BERT: Pre-training of Deep **Bidirectional** Transformers for Language Understanding*. NAACL 2019.

BERT: Pre-training of Deep Bidirectional Transformers for ...

Oct 11, 2018 — Unlike recent language representation models, **BERT** is designed to **pre-train deep bidirectional** representations from unlabeled text by jointly ...

by J Devlin · 2018 · Cited by 10265 · Related articles

Cloze Test (https://en.wikipedia.org/wiki/Cloze_test)

A **cloze test** (also **cloze deletion test**) is an exercise, test, or assessment consisting of a portion of language with certain items, words, or signs removed (cloze text), where the participant is asked to replace the missing language item. Cloze tests require the ability to understand context and vocabulary in order to identify the correct language or part of speech that belongs in the deleted passages. This exercise is commonly administered for the assessment of native and second language learning and instruction.

The word *cloze* is derived from *closure* in Gestalt theory. The exercise was first described by W.L. Taylor in 1953.[1]

A language teacher may give the following passage to students:

> Today, I went to the ________ and bought some milk and eggs. I knew it was going to rain, but I forgot to take my ________, and ended up getting wet on the way.

Students would then be required to fill in the blanks with words that would best complete the passage. Context in language and content terms is essential in most, if not all, cloze tests. The first blank is preceded by "the"; therefore, a noun, an adjective or an adverb must follow. However, a conjunction follows the blank; the sentence would not be grammatically correct if anything other than a noun were in the blank. The words "milk and eggs" are important for deciding which noun to put in the blank; "supermarket" is a possible answer; depending on the student, however, the first blank could be store, supermarket, shop, shops, market, or grocer while umbrella, brolly or raincoat could fit the second.

* Wilson L Taylor. 1953. *Cloze Procedure: A New Tool for Measuring Readability*. Journalism Bulletin, 30(4):415–433.

"Cloze Procedure": A New Tool for Measuring Readability ...

"**CLOZE PROCEDURE**" IS A **NEW** psychological **tool for measuring** the effectiveness of communication. The method is straightforward; the data are easily ...

by WL Taylor - 1953 - Cited by 2549 - Related articles

Cloze Test

Taylor (1953) first suggested the cloze procedure for determining the difficulty, or **"readability,"** of a text — reasoning that if several people could **reproduce the missing words of a "mutilated" passage** than the text must be easy to read, but if they could not supply the missing words the text must be difficult.

It was not long before Taylor himself saw the potential to use **cloze as a measure of reading comprehension**. In a 1956 paper, he reasoned that **"if the statement that a passage is 'readable' means that it is 'understandable,' then the scores that measure readability should measure comprehension too"**.

Further research has demonstrated that **cloze tests tend to correlate well with almost every kind of language test. Traditional cloze tests** were found to correlate not only with **reading comprehension** tests but also with **dictation**, **listening comprehension**, **structure/grammar**, and **vocabulary tests**.

Results such as these are what led Oller and Conrad (1971) and others to mark **cloze procedures as integrative tests and as good measures of *general language proficiency*** rather than the more specific reading comprehension alone.

* TREELA MCKAMEY. 2014. *Getting Closure On Cloze: A Validation Study of the "Rational Deletion" Method*.

Gestalt Laws of Grouping: Closure

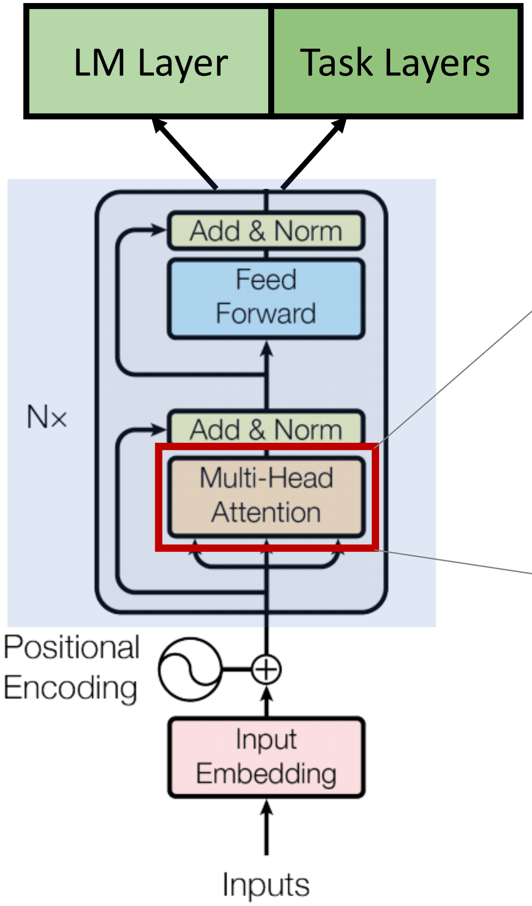
Gestalt psychologists emphasized that organisms perceive entire patterns or configurations, not merely individual components. The view is sometimes summarized using the adage, "**the whole is more than the sum of its parts**."

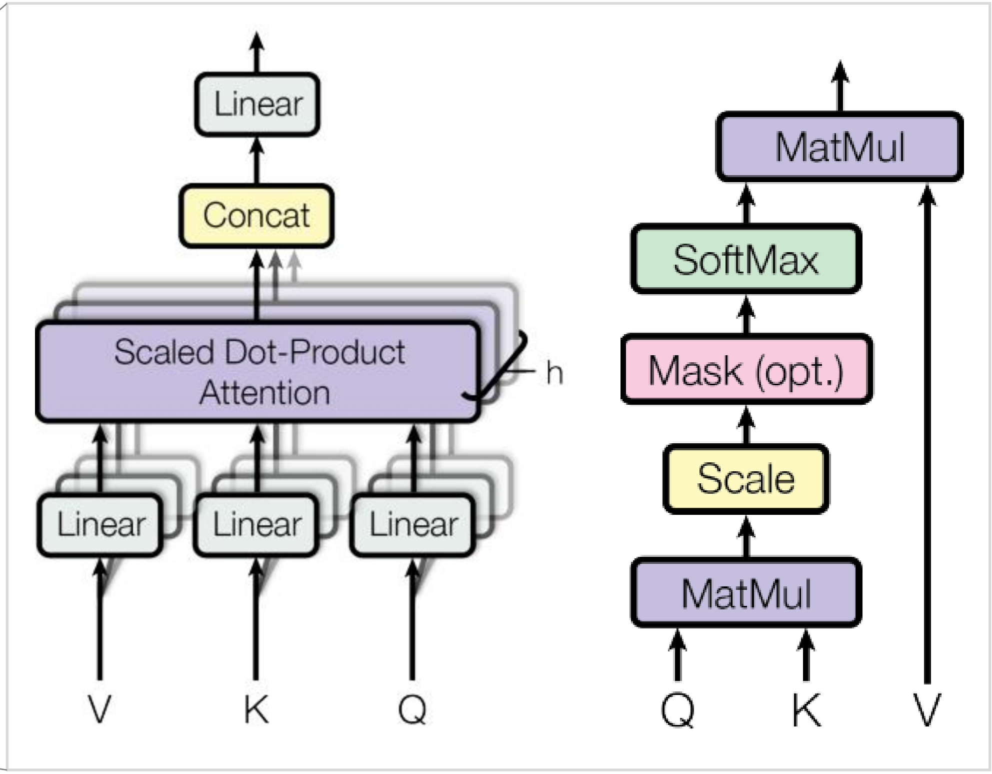
Gestalt psychologists believed that **humans tend to perceive objects as complete rather than focusing on the gaps that the object might contain**. For example, a circle has good Gestalt in terms of completeness. However, we will also perceive an incomplete circle as a complete circle. **That tendency to complete shapes and figures is called closure**. The law of closure states that individuals perceive objects such as shapes, letters, pictures, etc., as being whole when they are not complete. Specifically, **when parts of a whole picture are missing, our perception fills in the visual gap**.

* https://en.wikipedia.org/wiki/Gestalt_psychology

#2 - Backbone Networks: Multilayer Transformers

Relative position, self-attention, …





Multi-Head Attention

Scaled Dot-Product Attention

$$\mathrm{MultiHead}(Q,K,V) = \mathrm{Concat}(\mathrm{head_1},...,\mathrm{head_h})W^O$$
$$\text{where } \mathrm{head_i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Future Development of (LM) Pre-training

- **Pre-training tasks**
  - Parameterized self-supervised tasks
- **Network structures**
  - Improved (e.g. better/fast) Transformer & beyond
- Learning efficiency
  - Data/sample selection and planning (curriculum learning)
- Big & small pre-trained models
  - Large-scale training & optimization
  - The “ceiling” of small models

NMT & Pre-training

Pre-training for MT

Can we beat the BT baseline?
Low-resource?

|

MT for Pre-training

Can we have better pre-training tasks besides TLM/XLCo?
Can BT be used for multilingual LM pre-training?

Thanks