NLP Project

Neural Machine Translation

AI VIET NAM
Nguyen Quoc Thai

Outline

- Introduction
- NMT using Transformer
- NMT using Pre-trained LMs

Introduction

Translate a sentence $w^{(s)}$ in a **source language (input)** to a sentence $w^{(t)}$ in the **target language (output)**

Introduction

Translate a sentence $w^{(s)}$ in a **source language (input)** to a sentence $w^{(t)}$ in the **target language (output)**

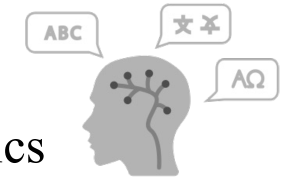
Automatic Speech Recognition (ASR)

translation of spoken language into text

Natural Language Understanding (NLU)

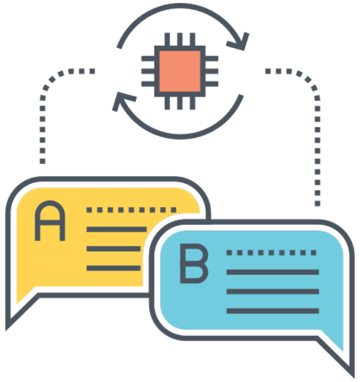
a computer's ability to understand language

- ❑ Syntax
- ❑ Semantics
- ❑ Phonology
- ❑ Pragmatics
- ❑ Morphology



Natural Language Generation (NLG)

generate natural language by a computer



Introduction

Translate a sentence $w^{(s)}$ in a **source language (input)** to a sentence $w^{(t)}$ in the **target language (output)**

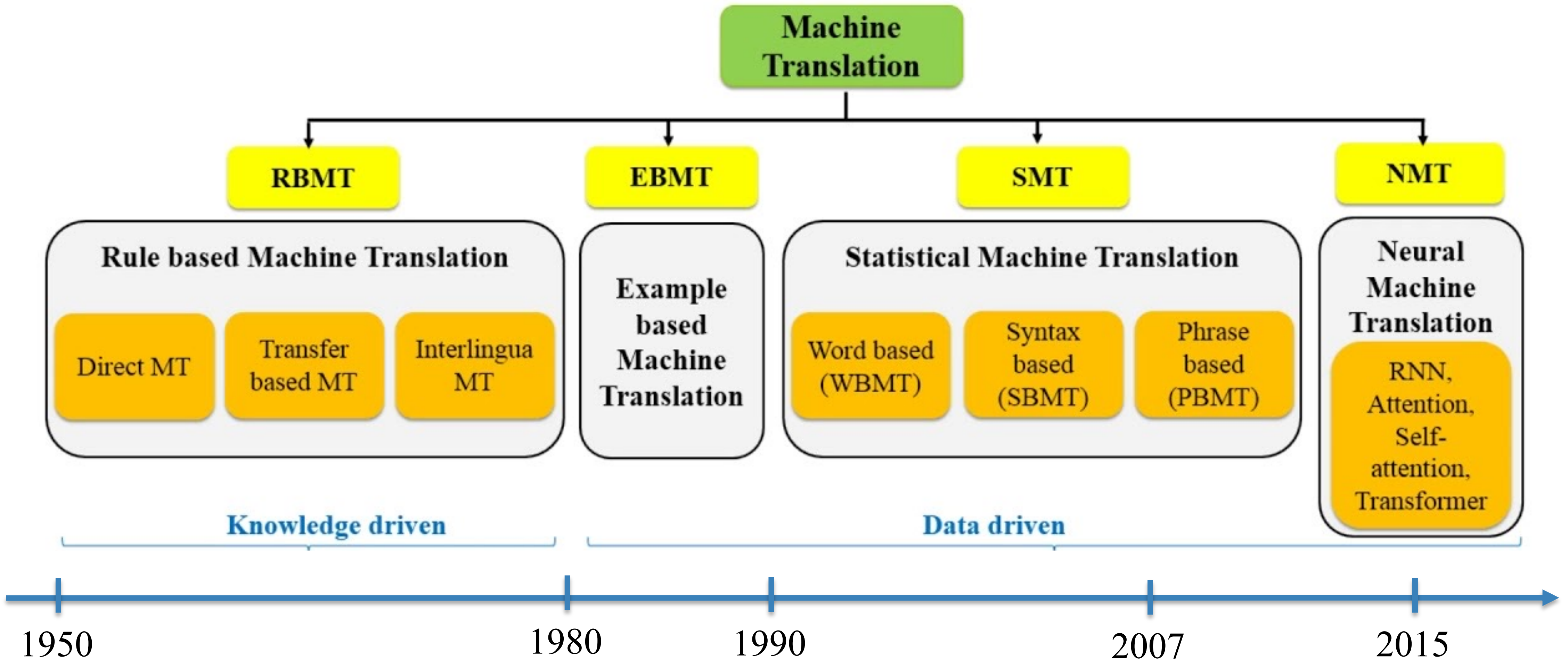
- Can be formulated as an optimization problem:

$$\hat{w}^{(t)} = \underset{w^{(t)}}{\operatorname{argmax}}\, \theta(w^{(s)}, w^{(t)})$$

Where θ is a scoring function over source and target sentences

- Requires two components:
  - Learning algorithm to compute parameters of θ
  - Decoding algorithm for computing the best translation $\hat{w}^{(t)}$

Introduction



Introduction

Evaluating translation quality

- Human judgement
  - Given: machine translation output
  - Given: source / reference translation
  - Task: assess the quality of machine translation output
- Different translations of "A Vinay le gusta Python"

To Vinay it like Python
Vinay debugs memory leaks
Vinay likes Python

Introduction

Evaluating translation quality

- Two main criteria:
  - Adequacy: Translation $w^{(t)}$ should adequately reflect the linguistic content of $w^{(s)}$
  - Fluency: Translation $w^{(t)}$ should be fluent text in the target language

- Different translations of "A Vinay le gusta Python"

	Adequate?	Fluent?
To Vinay it like Python	yes	no
Vinay debugs memory leaks	no	yes
Vinay likes Python	yes	yes

Introduction

Evaluating translation quality

- Two main criteria:
  - Adequacy: Translation $w^{(t)}$ should adequately reflect the linguistic content of $w^{(s)}$
  - Fluency: Translation $w^{(t)}$ should be fluent text in the target language

- Adequacy and fluency:

Adequacy	
5	All meaning
4	Most meaning
3	Much meaning
2	Little meaning
1	None

Fluency	
5	Flawless English
4	Good English
3	Non-native English
2	Disfluent English
1	Incomprehensible

Introduction

Evaluating Metrics

- Manual evaluation is most accurate, but expensive
- Automated evaluation metrics:
  - Compare system hypothesis with reference translations
  - BLEU Score (BiLingual Evaluation Understudy): Modified n-gram Precision
  - SacreBLEU Score (A Call for Clarity in Reporting BLEU Scores)

Introduction

Evaluating Metrics

Precision and Recall of words

System A	A	officials	~~responsibility~~	~~of~~	airport	~~safety~~	
Reference	A	officials	are	responsible	for	airport	security

- Precision:

$$\frac{\text{correct}}{\text{output} - \text{length}} = \frac{3}{6} = 50\%$$

- Recall:

$$\frac{\text{correct}}{\text{reference} - \text{length}} = \frac{3}{7} = 43\%$$

- F-measure:

$$\frac{\text{P x R}}{(\text{P} + \text{R})/2} = \frac{0.5 \text{ x } 0.43}{(0.5 + 0.43)/2} = 46\%$$

Introduction

Evaluating Metrics

Precision and Recall of words

- Flaw: no penalty for reordering

System A	A	officials	~~responsibility~~	~~of~~	airport	~~safety~~	
Reference	A	officials	are	responsible	for	airport	security
System B	airport	security	A	officials	are	responsible	

Metric	System A	System B
Precision	50%	100%
Recall	43%	86%
F-measure	46%	92,5%

Introduction

Evaluating Metrics

BLEU

- N-gram overlap between machine translation output and reference translation
- Compute precision for n-grams of size 1 to 4
- Add brevity penalty (for too short translations)

$$\text{BLEU} = \min\left(1, \frac{\text{output} - \text{length}}{\text{reference} - \text{length}}\right)\left(\prod_{i=1}^{4} \text{precision}_1\right)^{1/4}$$

- Typically computed over the entire corpus, not single sentences

AI VIET NAM
@aivietnam.edu.vn

Introduction

Evaluating Metrics

BLEU

1-gram

System A	A	officials	~~responsibility~~	~~of~~	$\boxed{\text{airport}}$	~~safety~~	
Reference	A	officials	are	responsible	for	airport	security
System B	airport	security	A	officials	are	responsible	

Metric	System A	System B
Precision (1 gram)	3/6	6/6
Precision (2 gram)		
Precision (3 gram)		
Precision (4 gram)		
Brevity penalty		
BLEU		

Introduction

Evaluating Metrics

BLEU

System A	A	officials	~~responsibility~~	~~of~~	~~airport~~	~~safety~~	
Reference	A	officials	are	responsible	for	airport	security
System B	airport	security	A	officials	are	responsible	

2 -gram

Metric	System A	System B
Precision (1 gram)	3/6	6/6
Precision (2 gram)	1/5	4/5
Precision (3 gram)	0/4	2/4
Precision (4 gram)	0/3	1/4
Brevity penalty	6/7	6/7
BLEU	0	0.52

Introduction

Evaluating Metrics

BLEU

$$\log BLEU = \min\left(1 - \frac{r}{c}, 0\right) + \sum_{n=1}^{N} w_n \log p_n$$

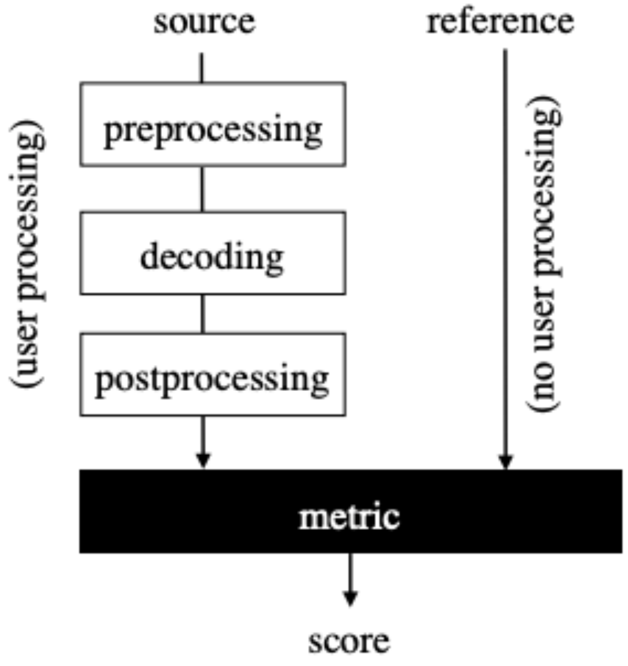
r: reference-length, c: output (candidate)-length

n: n-gram (1,2,3,4), w_n: weight of n-gram

uniform weights $w_n=1/n$

p_n: precision n-gram

SacreBLEU (A Call for Clarity in Reporting BLEU)

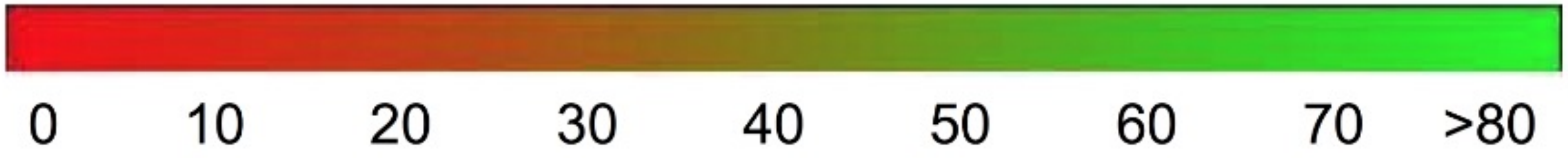


Introduction

Evaluating Metrics

BLEU Score	Interpretation
< 10	Almost useless
10 - 19	Hard to get the gist
20 - 29	The gist is clear, but has significant grammatical errors
30 - 40	Understandable to good translations
40 - 50	High quality translations
50 - 60	Very high quality, adequate, and fluent translations
> 60	Quality often better than human

The following color gradient can be used as a general scale interpretation of the BLEU score:



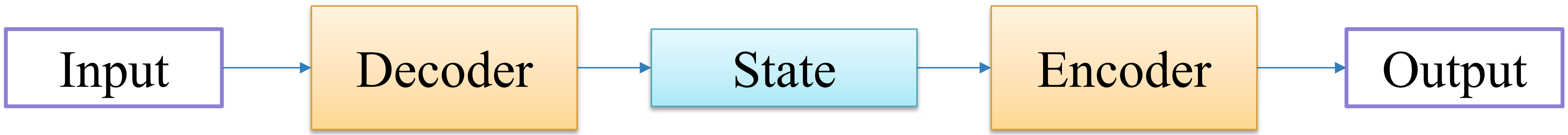
Outline

- Introduction
- NMT using Transformer
- NMT using Pre-trained LMs

NMT using Transformer

Sequence to Sequence

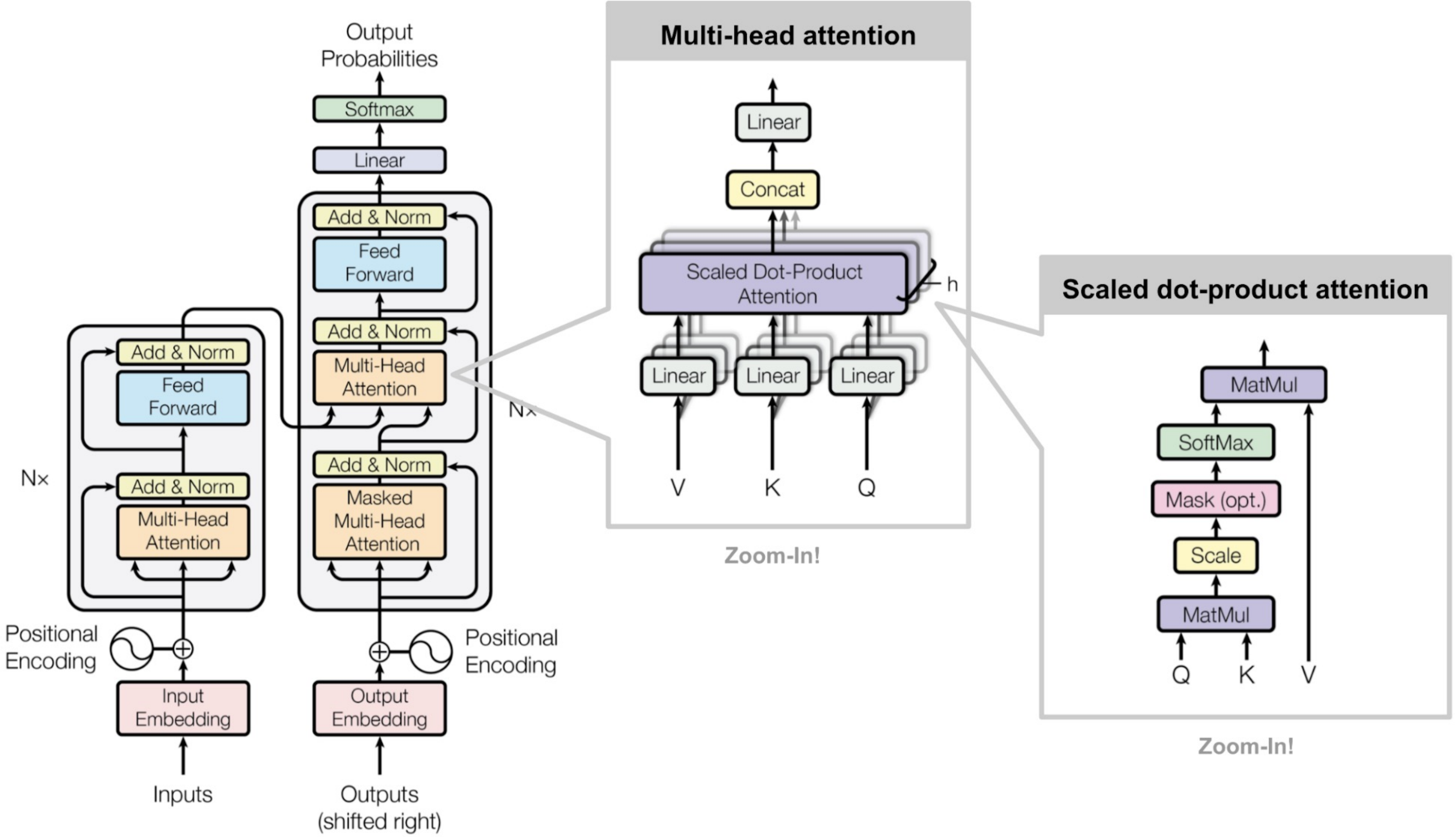
- A single neural network is used to translate from source to target
- Architecture: Encoder-Decoder
- Encoder: Convert source sentence (input) into a vector/matrix (State)
- Decoder: Convert encoding into a sentence in target language (output)



Thought Vector
Capture all information of input sentence

NMT using Transformer

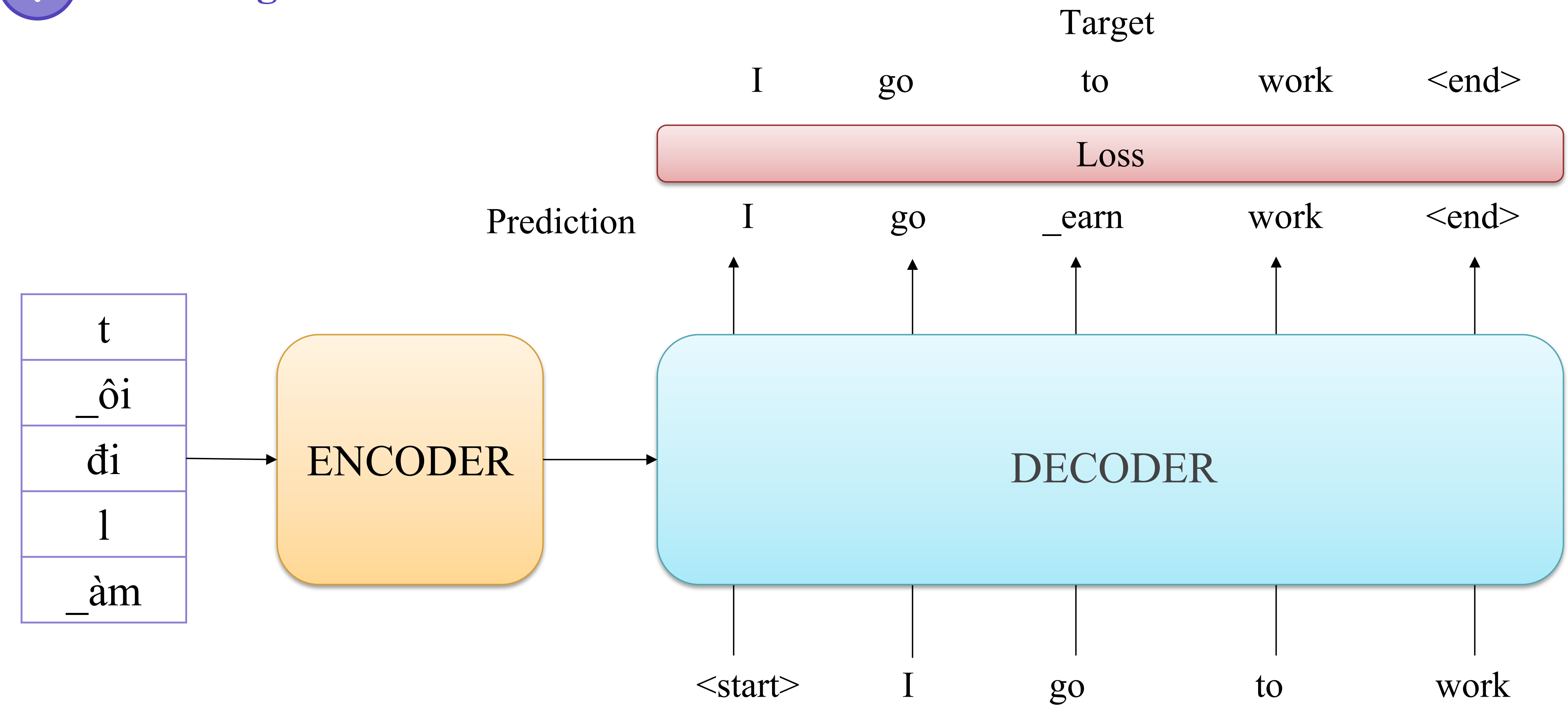
Transformer Model



AI VIET NAM
@aivietnam.edu.vn

NMT using Transformer

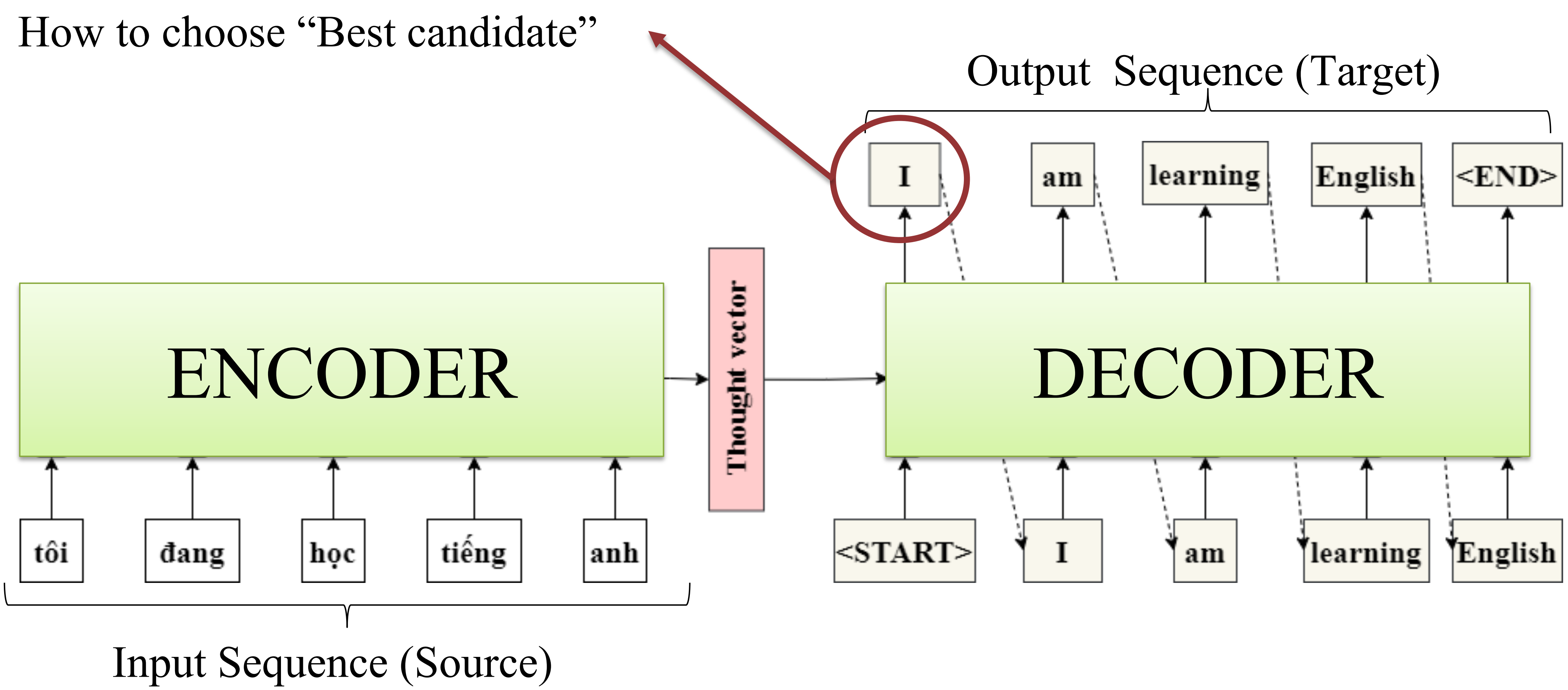
Training



AI VIET NAM
@aivietnam.edu.vn

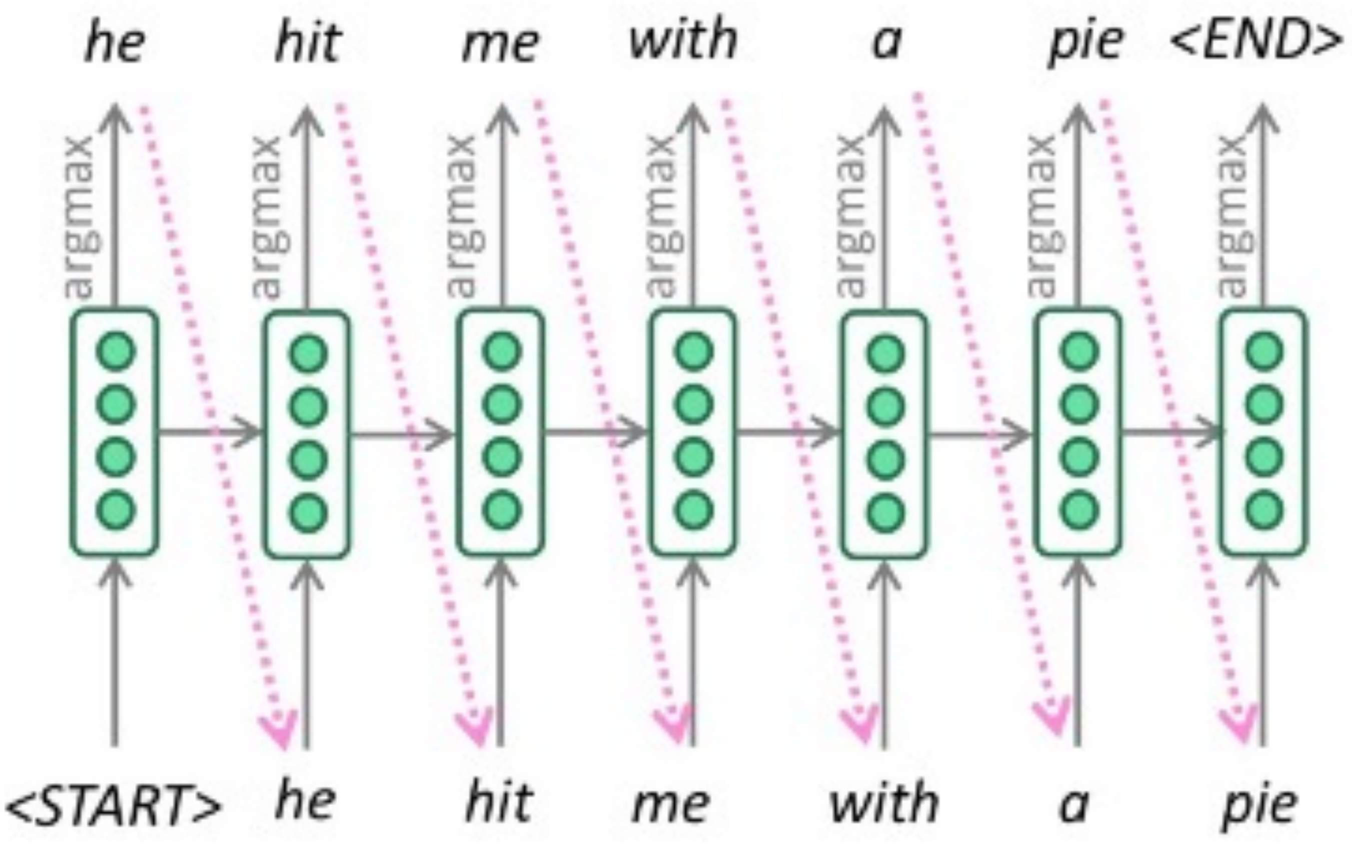
NMT using Transformer

Training



NMT using Transformer

Greedy Decoding

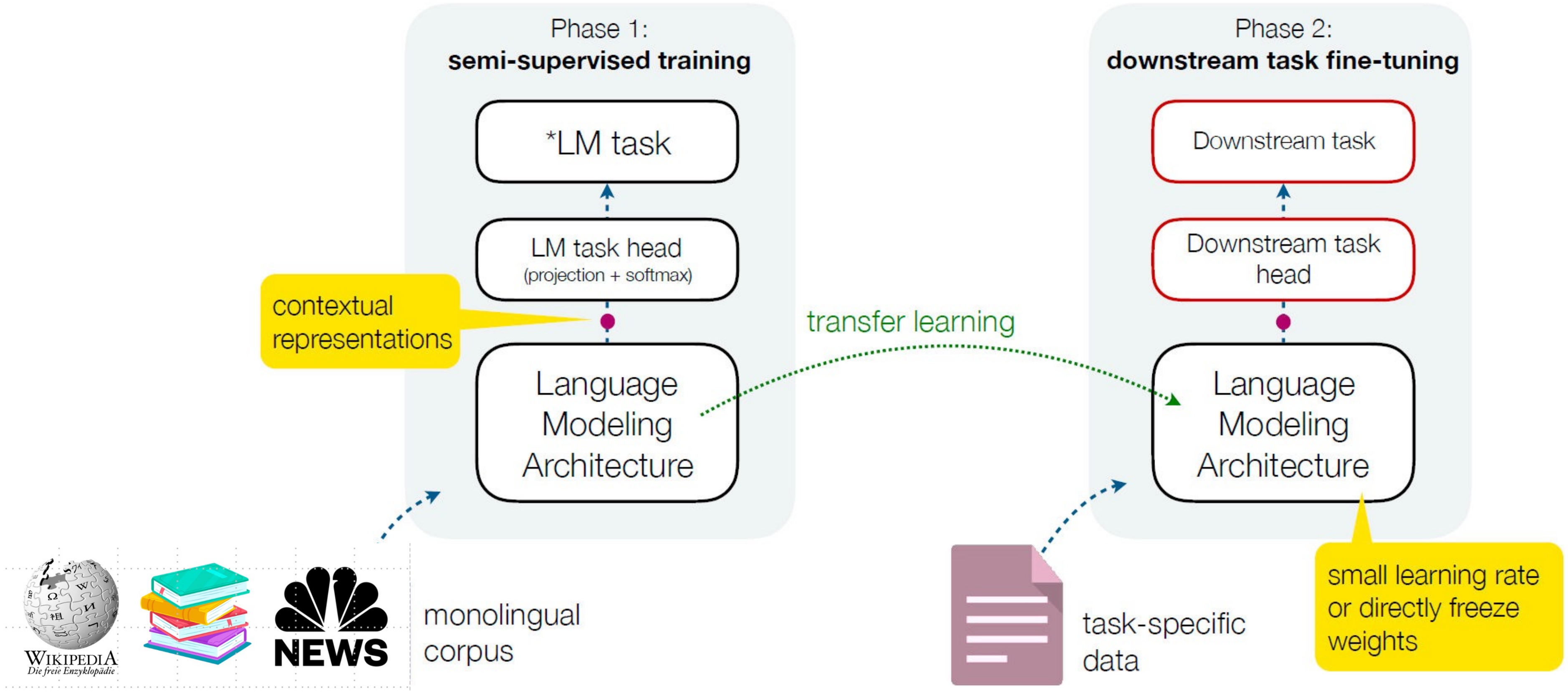


Outline

- Introduction
- NMT using Transformer
- NMT using Pre-trained LMs

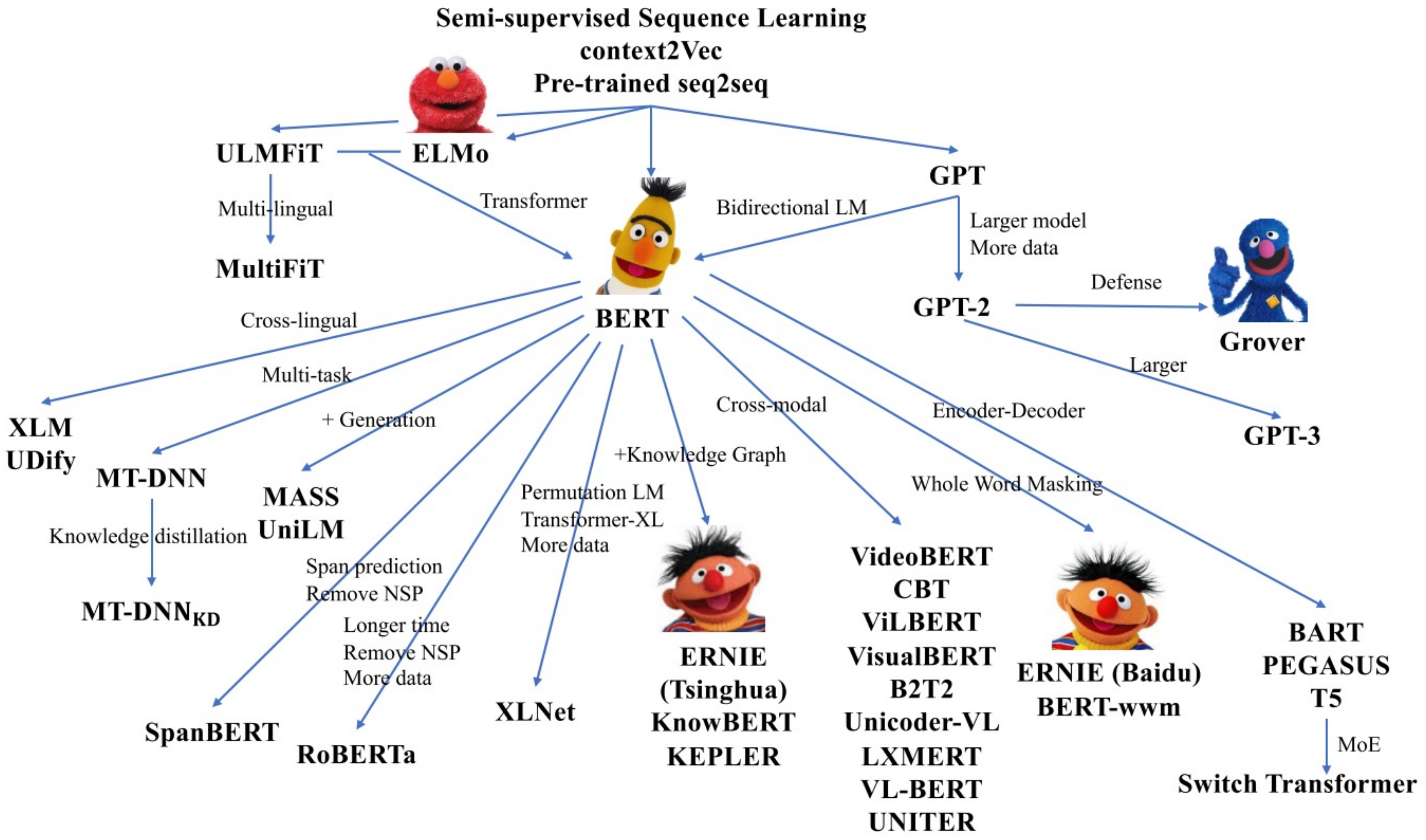
NMT using Pre-trained LMs

Pre-trained LMs



NMT using Pre-trained LMs

Pre-trained LMs

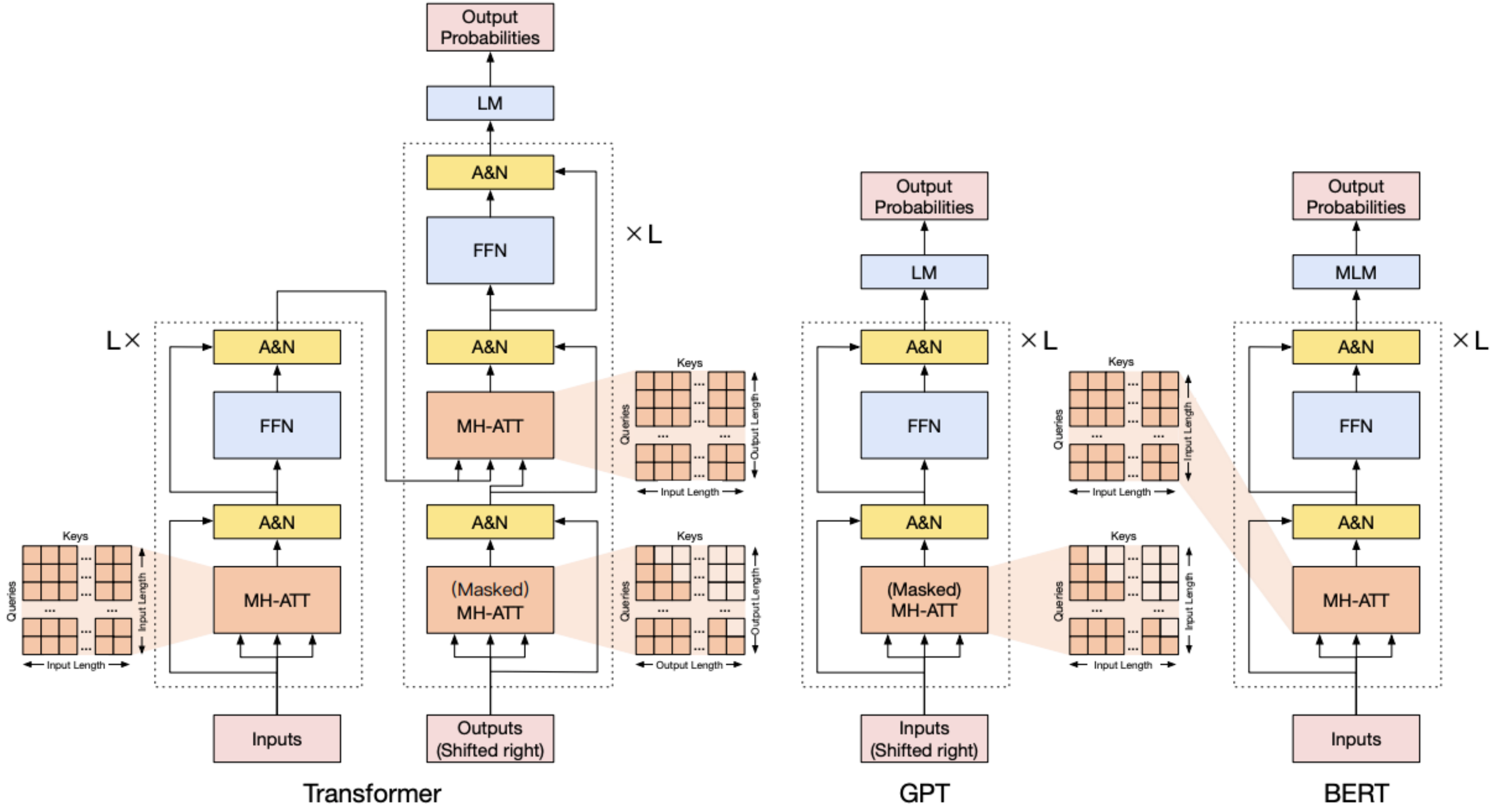


Source

AI VIET NAM
@aivietnam.edu.vn

NMT using Pre-trained LMs

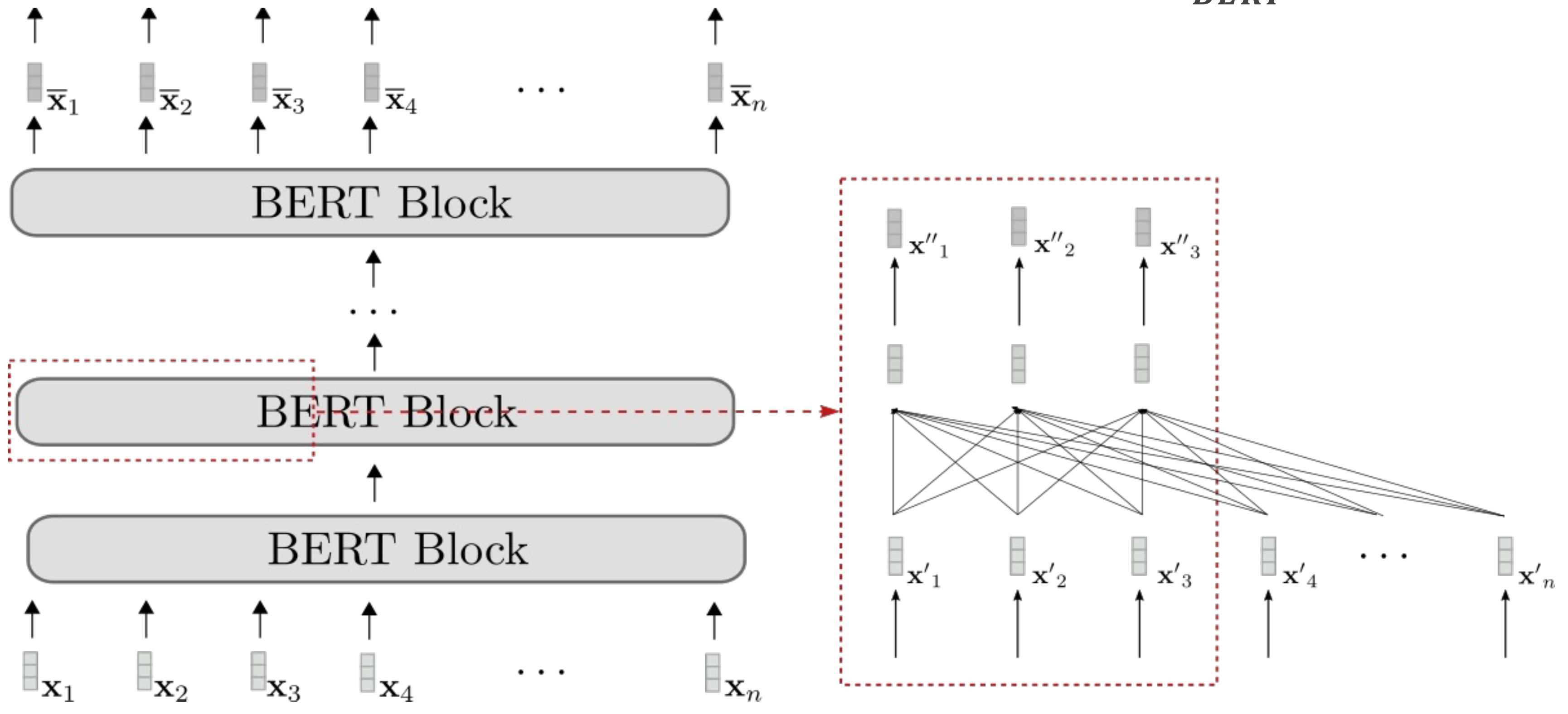
Pre-trained LMs



NMT using Pre-trained LMs

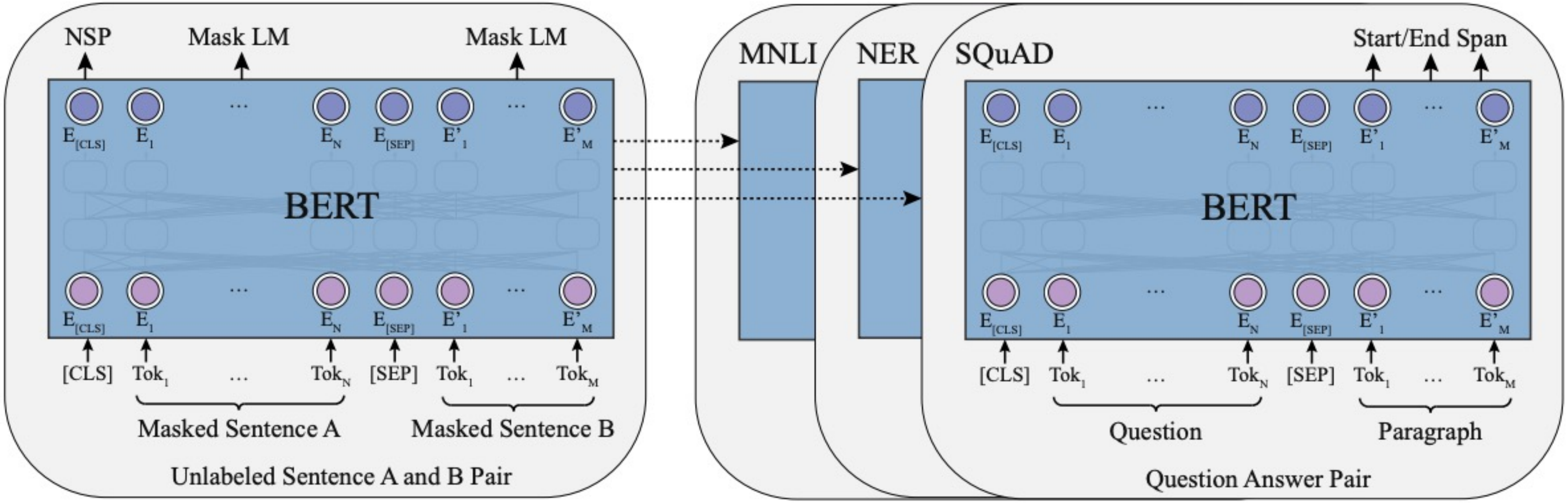
Pre-trained LMs: BERT

- BERT: An encoder-only model
- Maps an input sequence to a contextualized sequence: $f_{\theta_{BERT}}: X_{1:n} \longrightarrow \bar{X}_{1:n}$



NMT using Pre-trained LMs

Pre-trained LMs: BERT



Pre-training

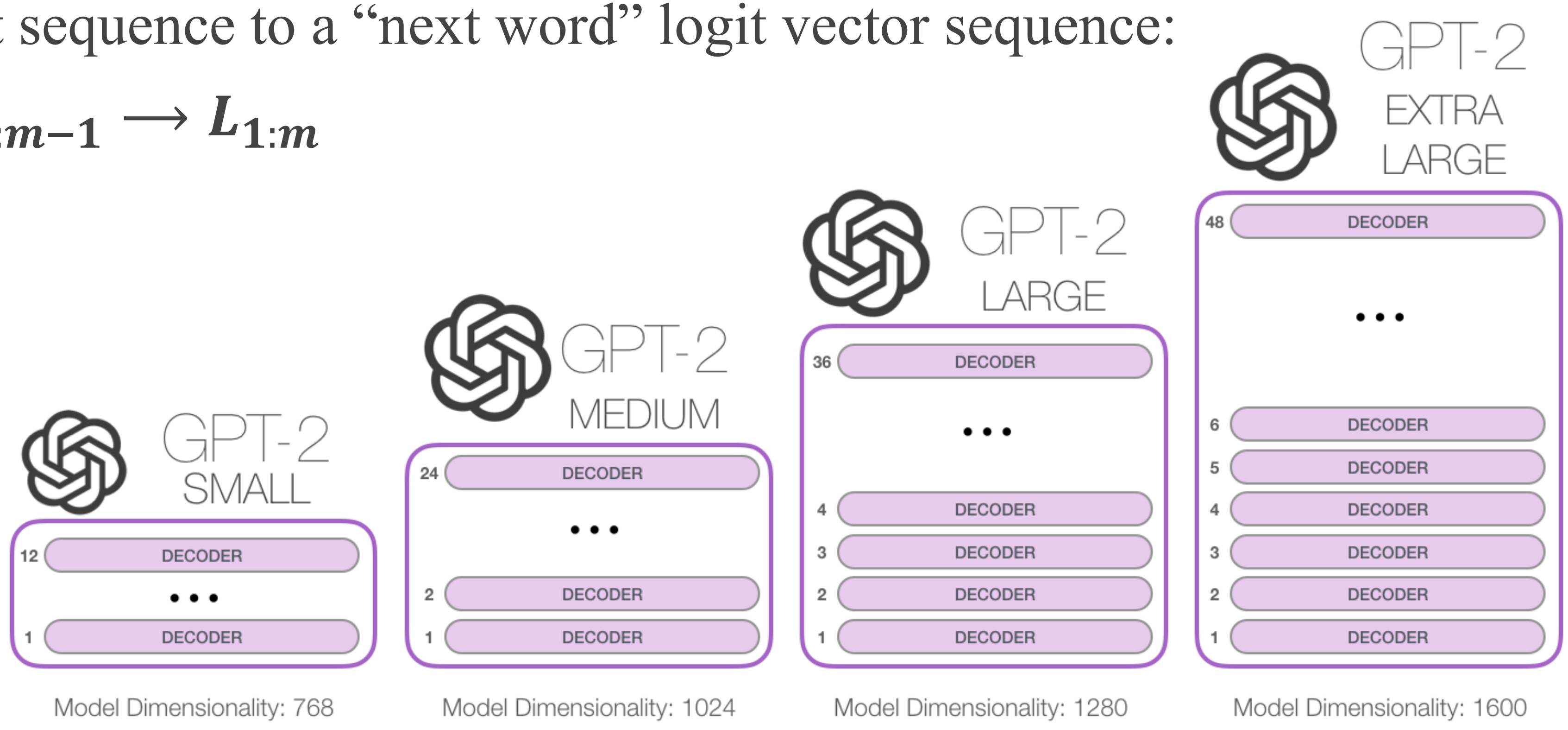
Fine-Tuning

NMT using Pre-trained LMs

Pre-trained LMs: GPT2

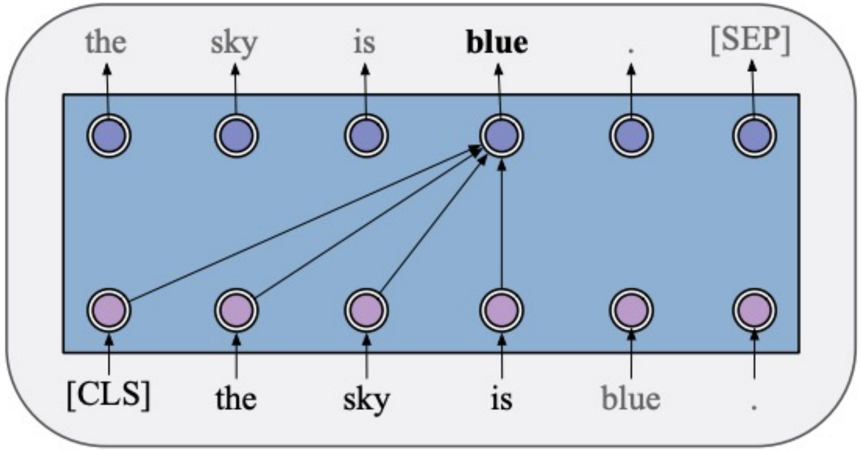
- GPT2: A decoder-only model, use uni-directional (causal) self-attention
- Maps an input sequence to a “next word” logit vector sequence:

$$f_{\theta_{GPY2}}: X_{0:m-1} \longrightarrow L_{1:m}$$

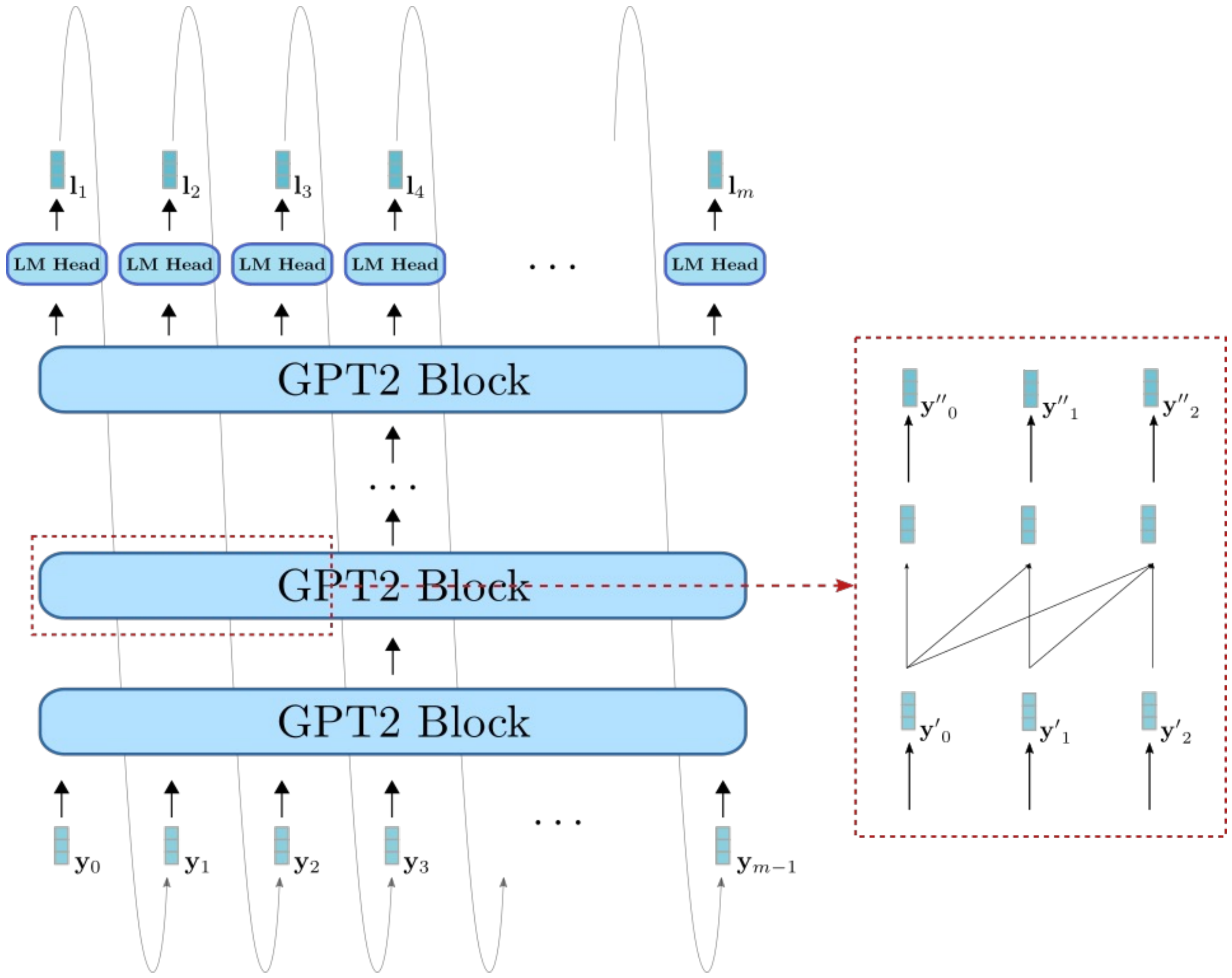


NMT using Pre-trained LMs

Pre-trained LMs: GPT2

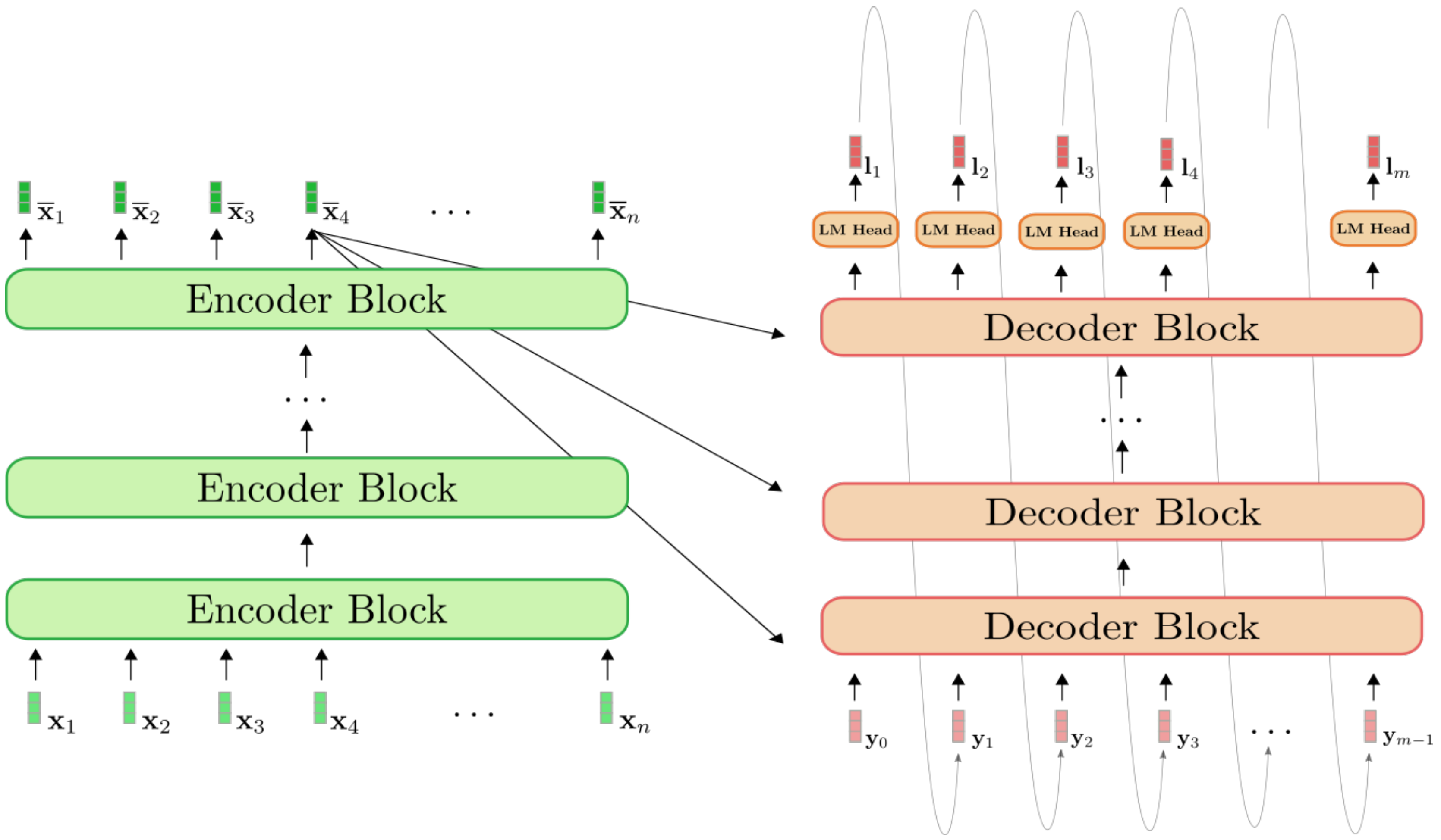


GPT



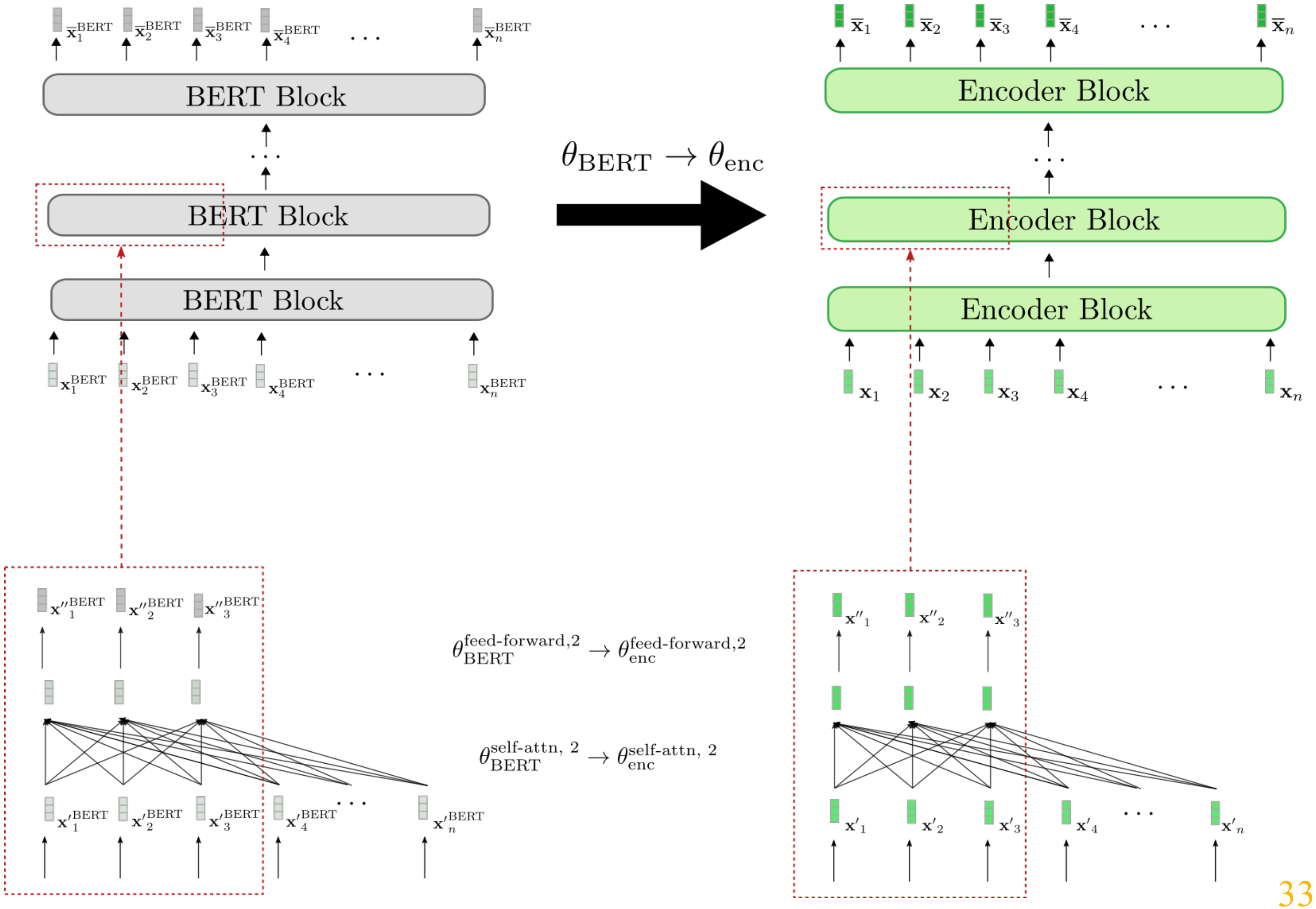
NMT using Pre-trained LMs

Encoder-Decoder with BERT and GPT2



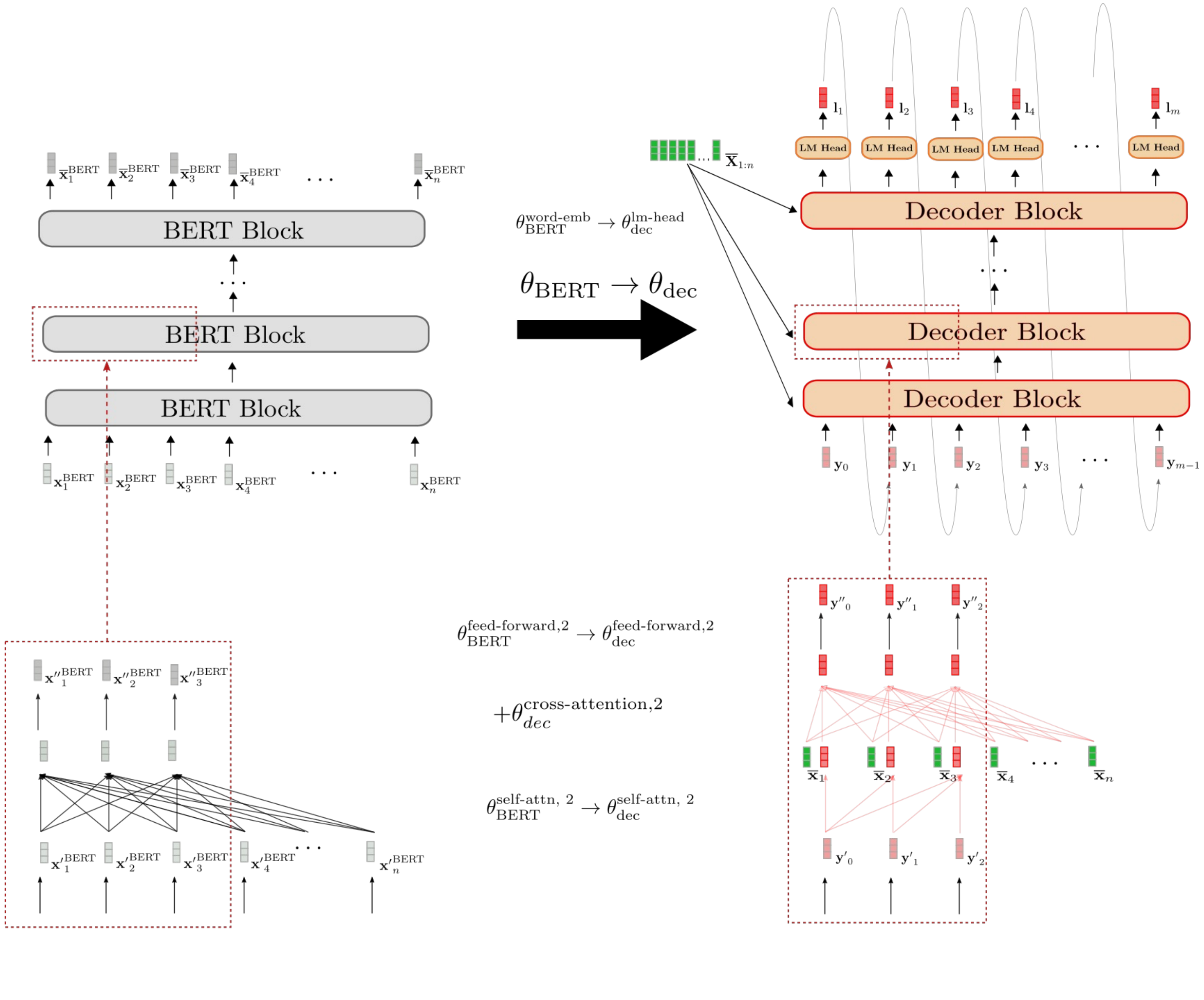
NMT using Pre-trained LMs

BERT for Encoder



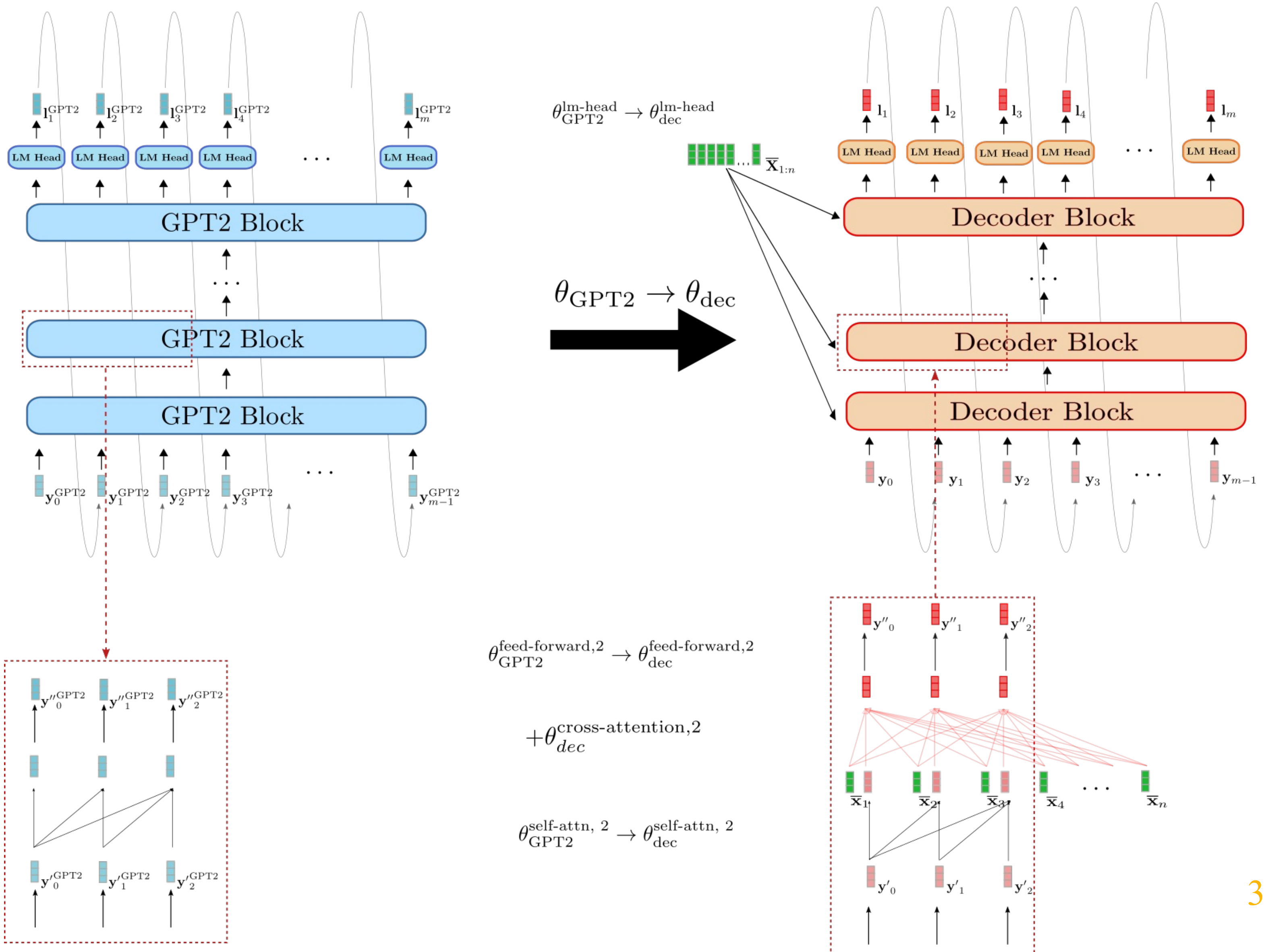
NMT using Pre-trained LMs

BERT for Decoder



NMT using Pre-trained LMs

GPT2 for Decoder



NMT using Pre-trained LMs

Experiment

- **Dataset: IWSLT'15 English-Vietnamese**

Training: 133 317 Validation: 1 553 Test: 1 269

Experiment	Model	ScareBLEU
#1	Standard Transformer (Greedy Search)	24.66 55.9/30.3/18.5/11.8
#2	BERT-to-BERT (Greedy Search)	25.41 53.8/31.8/19.8/12.3
#3	BERT-to-GPT2 (Greedy Search)	23.56 49.1/28.5/18.4/12.0

Thanks!

Any questions?