CS11-747 Neural Networks for NLP

Advanced Sequence-to-sequence Models

Graham Neubig

Carnegie Mellon University
Language Technologies Institute

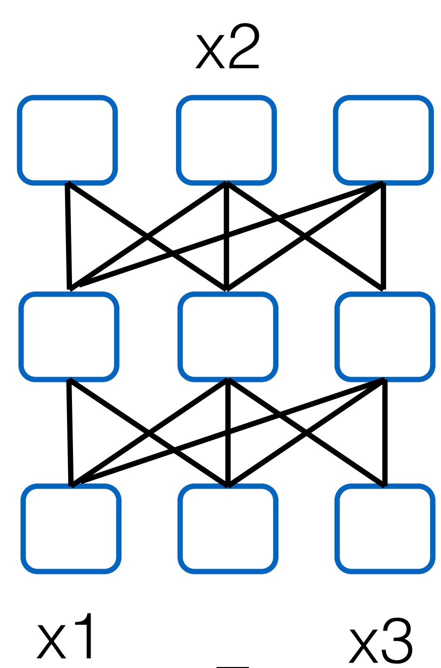
Site
https://phontron.com/class/nn4nlp2021/

Slides by Pengfei Liu

Remember: Masked Language Model

BERT, RoBERTa

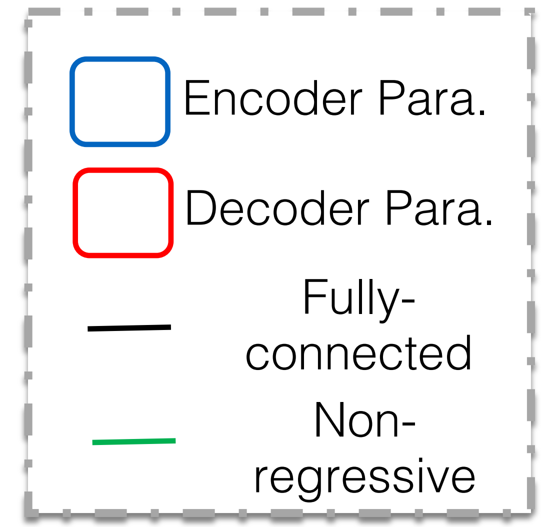
Masked language model (LM)

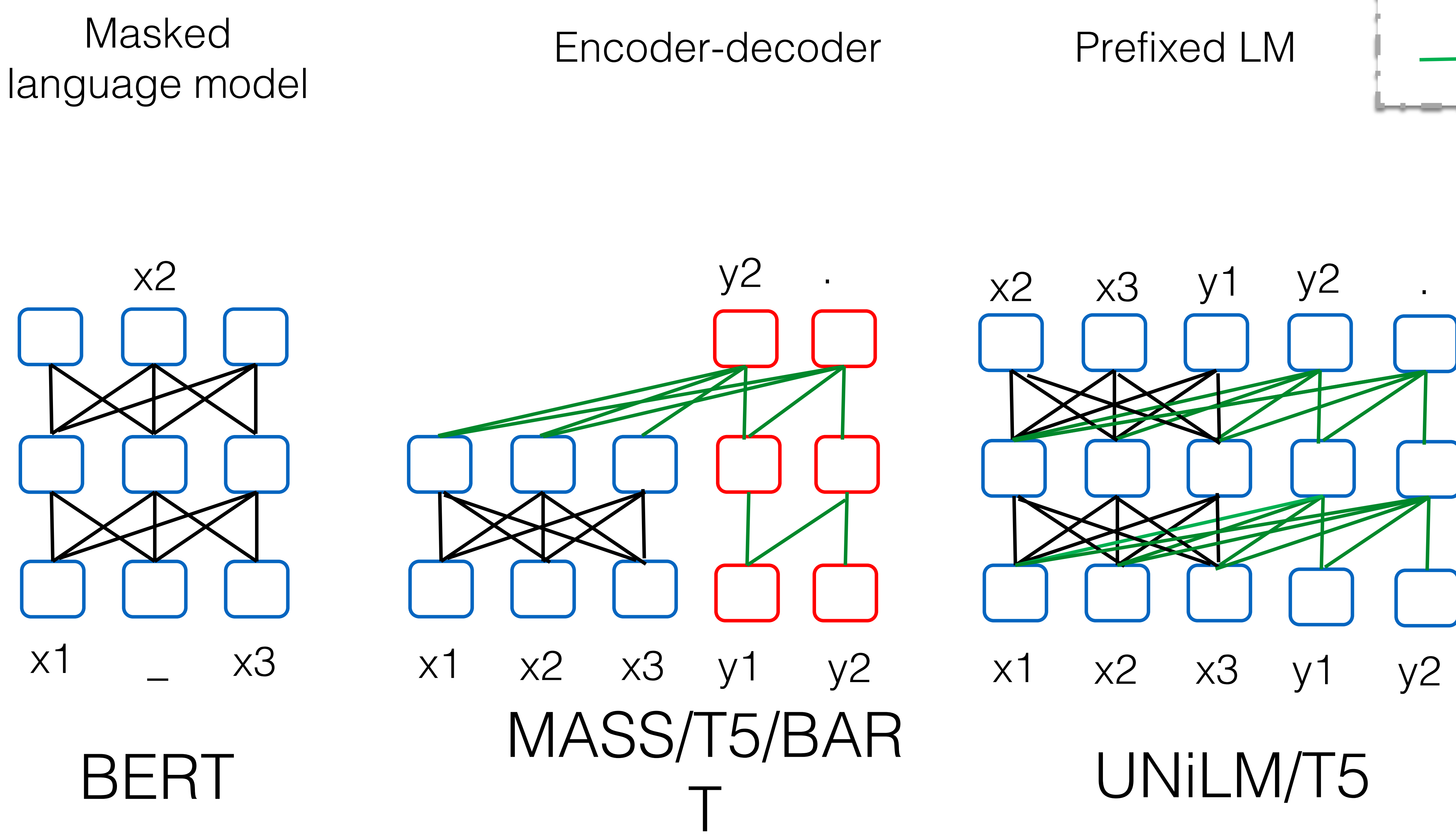


- Encoder-only pre-trained model
- Not suitable for conditional generation

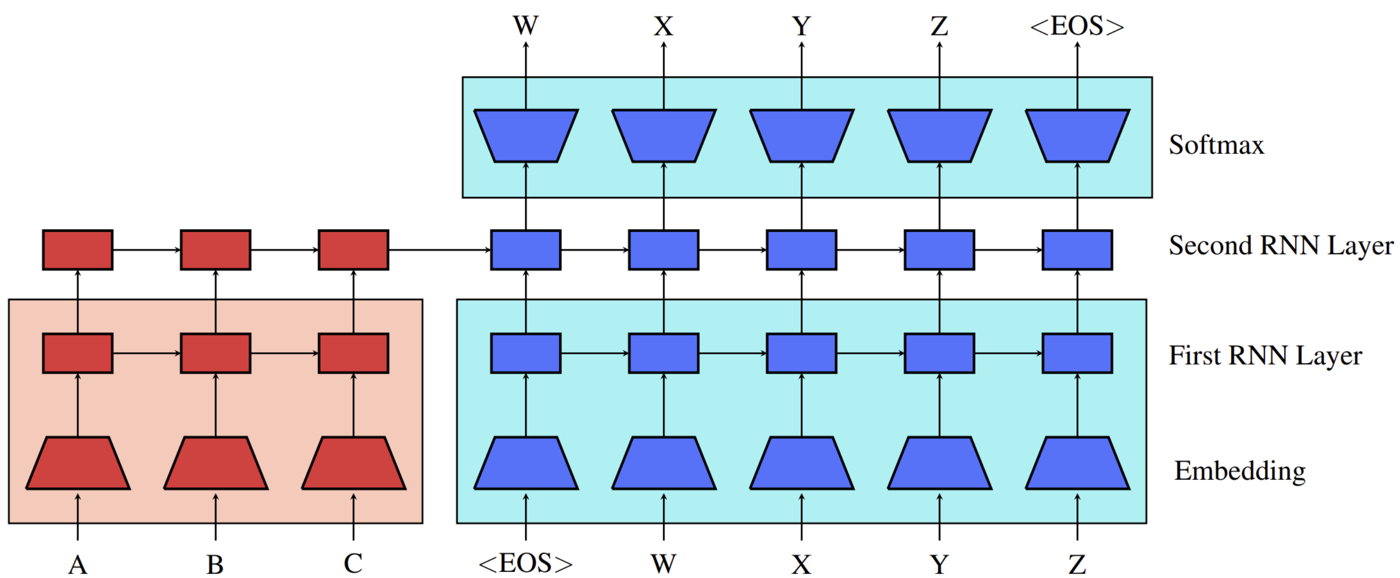
How to design a pre-trained model that can adapt to conditional generation?

Seq2seq Pretraining and Beyond





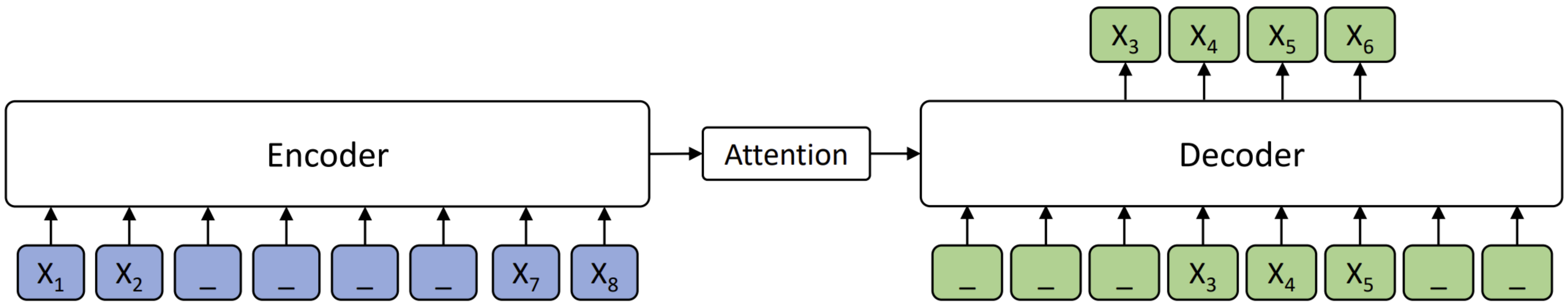
PreSeq2seq
(Ramachandran et al.)



- **Model:** RNN-based Encoder-decoder, no self-attention
- **Objective:** *language model*
  - *Encoder & decoder are pre-trained by two language models*
- **Data:** Task-specific

MASS

(Song et al.)

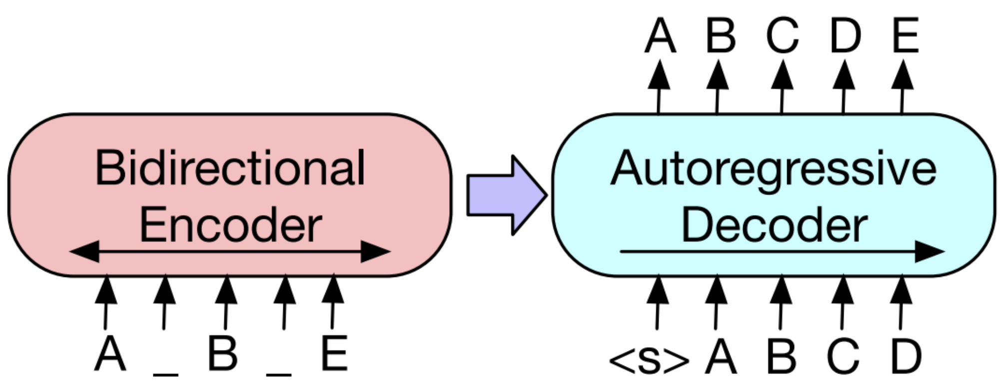


- **Model:** Transformer-based Encoder-decoder
- **Objective:** *only* predict masked spans
- **Data:** WebText

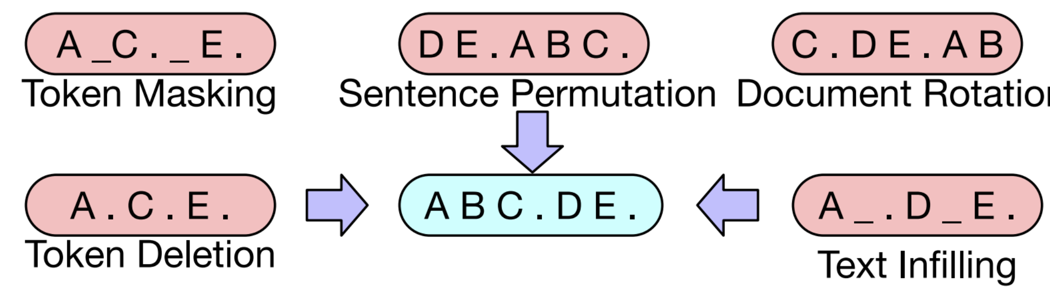
BART

(Lewis et al.)

Framework

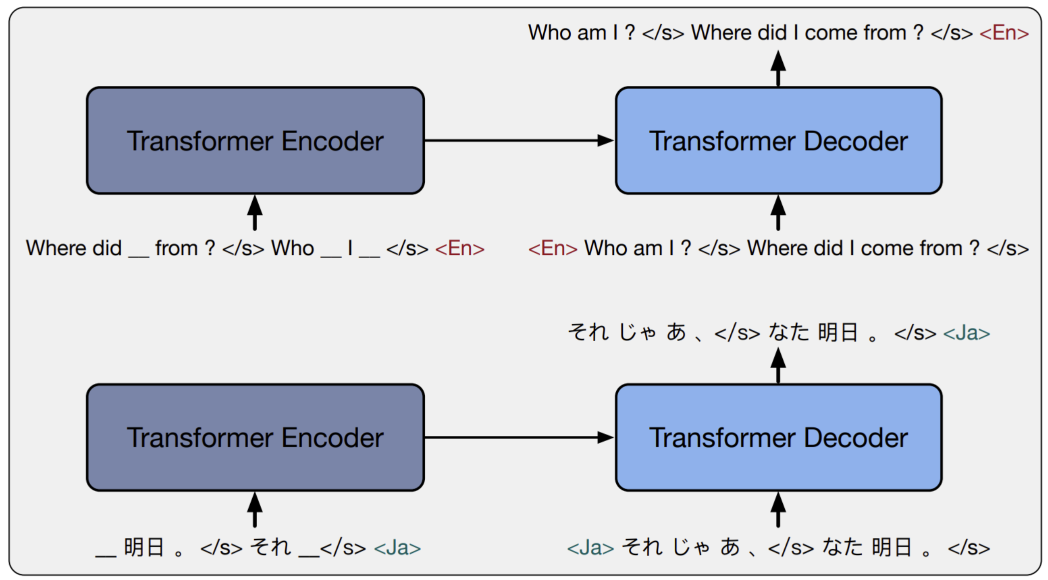


Different Corruption



- **Model:** Transformer-based encoder-decoder model
- **Objective:** Re-construct (corrupted) *original sentences*
- **Data:** similar to RoBERTa (160GB): BookCorpus, CC-NEWs, WebText, Stories

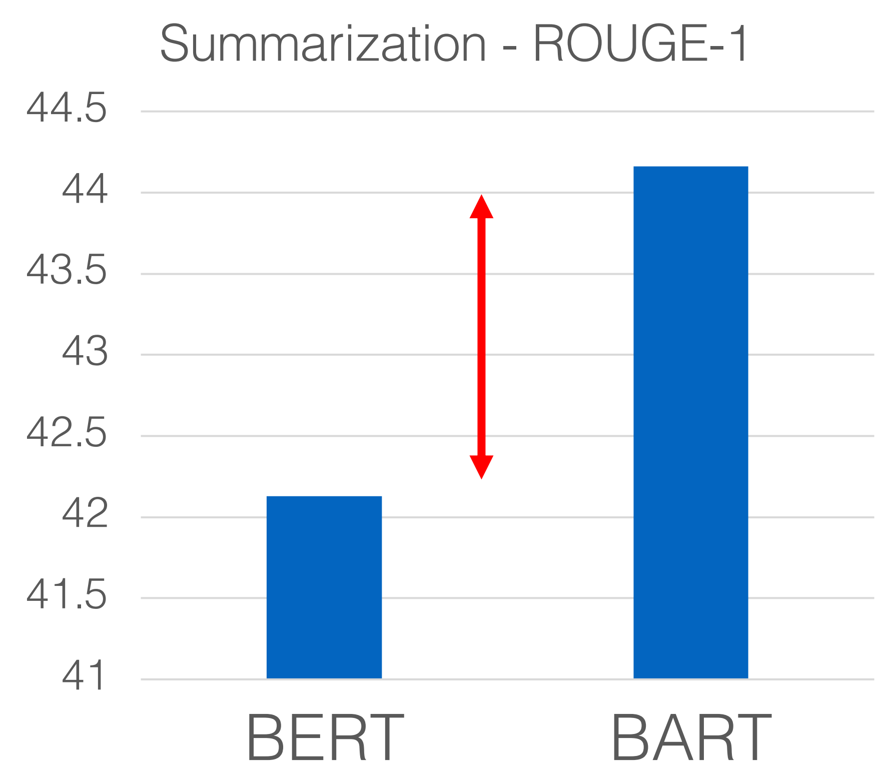
mBART (Liu et al.)



- **Model:** Transformer-based *Multi-lingual Denoising* auto-encode
- **Objective:** Re-construct (corrupted) *original sentences*
- **Data:** CC25 Corpus (25 langauges)

Seq2seq v.s Masked LM

- Regarding generation tasks:
  - Seq2seq (BART) could significantly outperform masked LM (BERT)



~ 2.0 ROUGE-1 is a fairly large performance gap

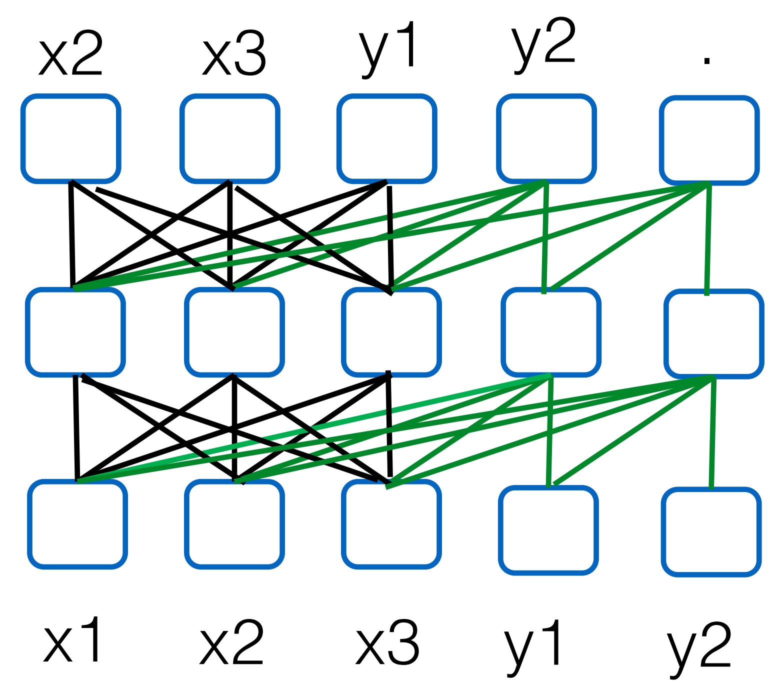
Seq2seq v.s Masked LM

- Regarding non-generation tasks:
  - Seq2seq (BART) could achieve comparable or even better performance than masked LM (BERT)

	SQuAD 1.1 EM/F1	SQuAD 2.0 EM/F1	MNLI m/mm	SST Acc	QQP Acc	QNLI Acc	STS-B Acc	RTE Acc	MRPC Acc	CoLA Mcc
BERT	84.1/90.9	79.0/81.8	86.6/-	93.2	91.3	92.3	90.0	70.4	88.0	60.6
UniLM	-/-	80.5/83.4	87.0/85.9	94.5	-	92.7	-	70.9	-	61.1
XLNet	**89.0**/94.5	86.1/88.8	89.8/-	95.6	91.8	93.9	91.8	83.8	89.2	63.6
RoBERTa	88.9/**94.6**	**86.5/89.4**	**90.2/90.2**	96.4	92.2	94.7	**92.4**	86.6	**90.9**	**68.0**
BART	88.8/**94.6**	86.1/89.2	89.9/90.1	**96.6**	**92.5**	**94.9**	91.2	**87.0**	90.4	62.8

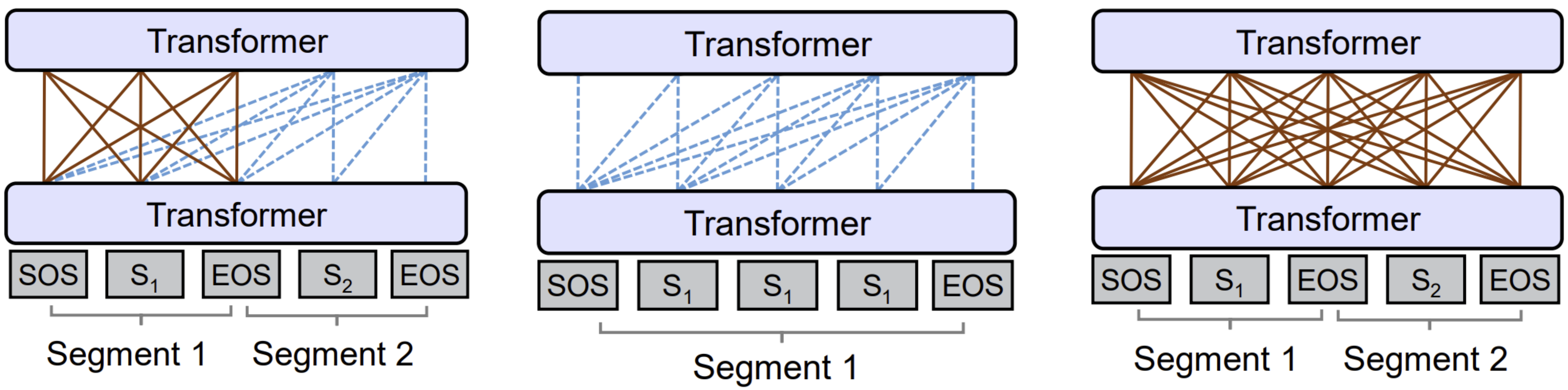
Prefixed Language Model

- Encoder and decoder are put in the same Transformer by using
  - Fully-connected self-attention
  - Left-to-right self-attention



UNiLM

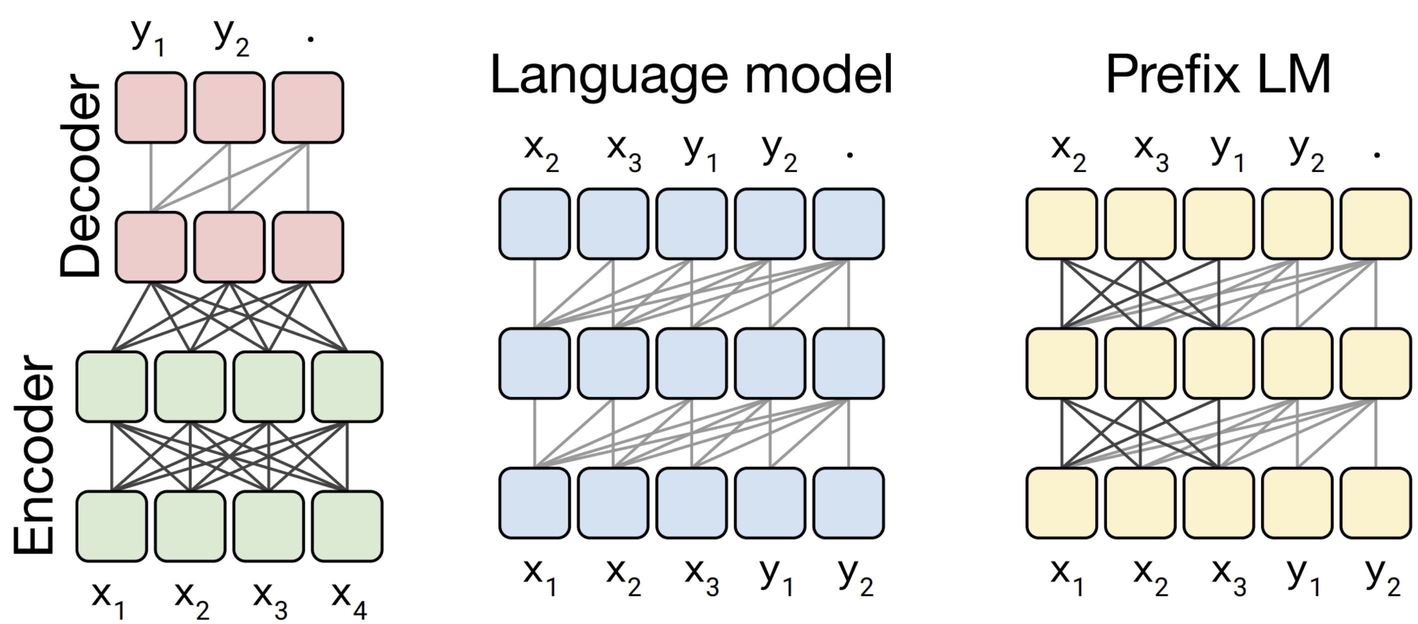
(Dong et al.)



- **Model:** prefixed-LM, left-to-right LM, Masked LM
- **Objective:** three types of LMs, *shared* parameters
- **Data:** English Wikipedia and BookCorpus

T5

(Raffel et al.)



- **Model:** left-to-right LM, Prefixed LM, encode-decoder
- **Objective:** explore different cases respectively
- **Data:** C4 (750G) + Wikipedia + RealNews + WebText

T5

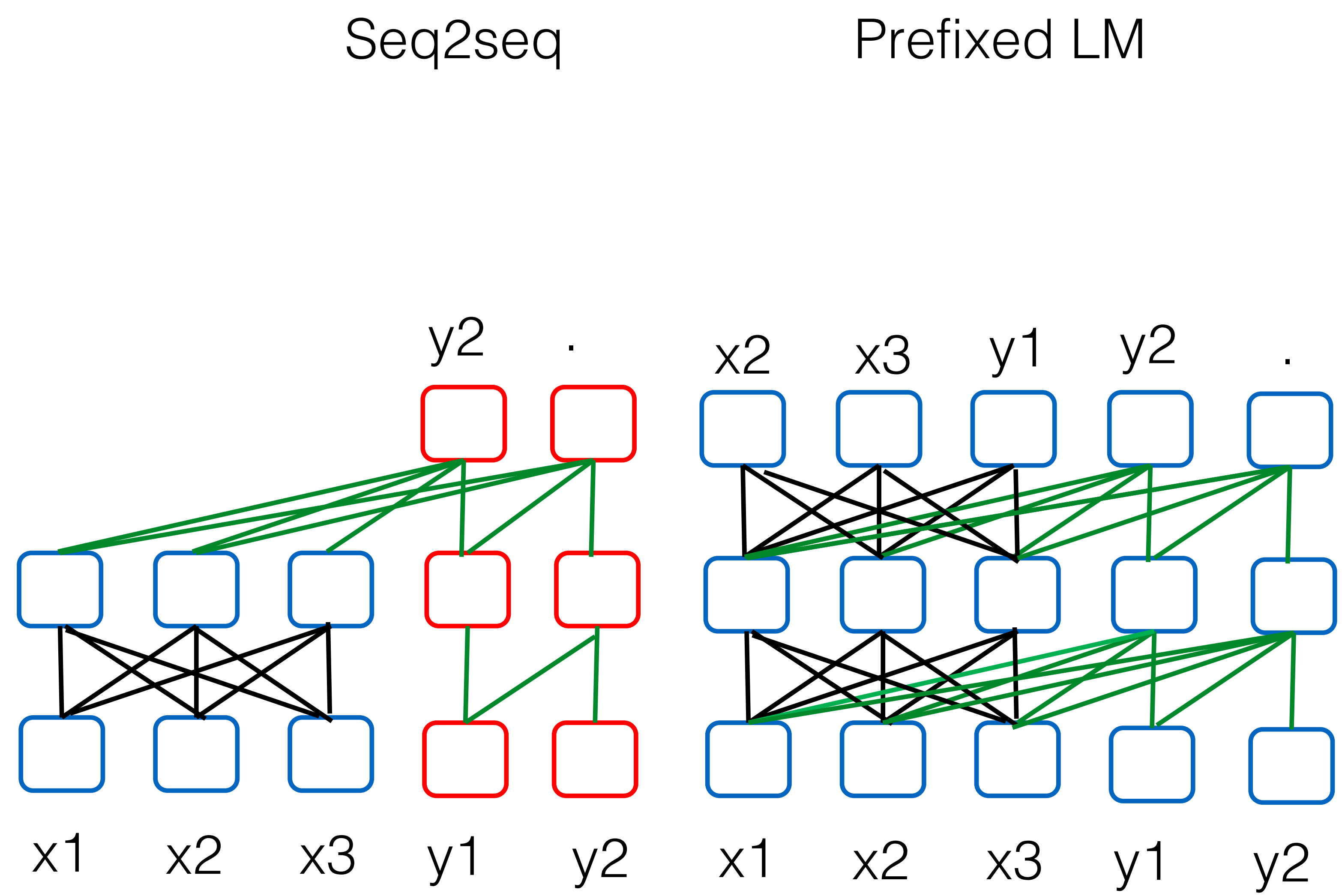
(Raffel et al.)

Objective	Inputs	Targets
Prefix language modeling	Thank you for inviting	me to your party last week .
BERT-style Devlin et al. (2018)	Thank you <M> <M> me to your party apple week .	*(original text)*
Deshuffling	party me for your to . last fun you inviting week Thank	*(original text)*
MASS-style Song et al. (2019)	Thank you <M> <M> me to your party <M> week .	*(original text)*
I.i.d. noise, replace spans	Thank you <X> me to your party <Y> week .	<X> for inviting <Y> last <Z>
I.i.d. noise, drop tokens	Thank you me to your party week .	for inviting last
Random spans	Thank you <X> to <Y> week .	<X> for inviting me <Y> your party last <Z>

- **Model:** left-to-right LM, Prefix LM, encode-decoder
- **Objective:** explore different cases respectively
- **Data:** C4 (750G) + Wikipedia + RealNews + WebText

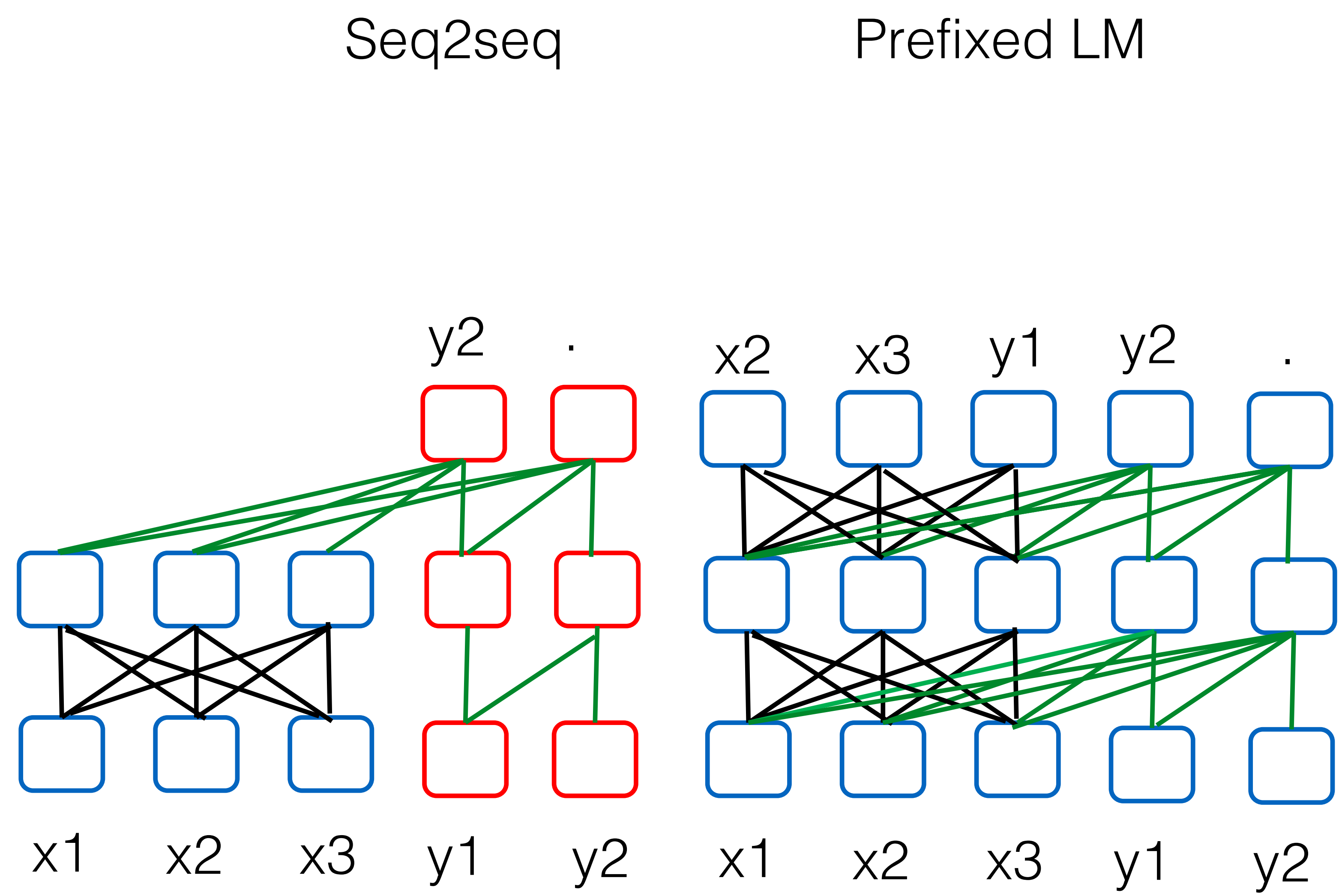
Seq2seq v.s Prefixed LM

- Architecture
  - Seq2seq: different parameters for encoder & decoder
  - Prefixed LM: same parameters for prefix and continuation.



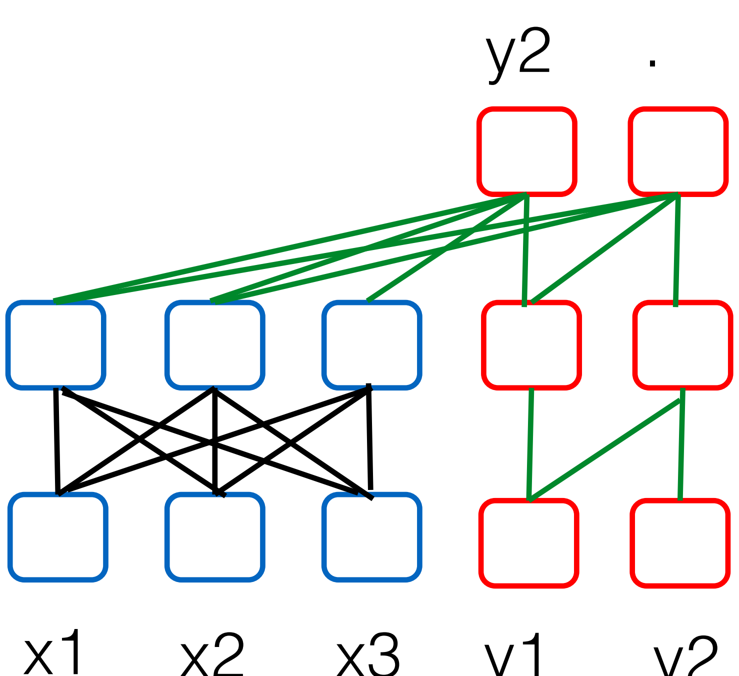
Seq2seq v.s Prefixed LM

- Loss Function
  - Seq2seq: calculate loss only on output
  - Prefixed LM: calculate loss on both prefix and continuation.



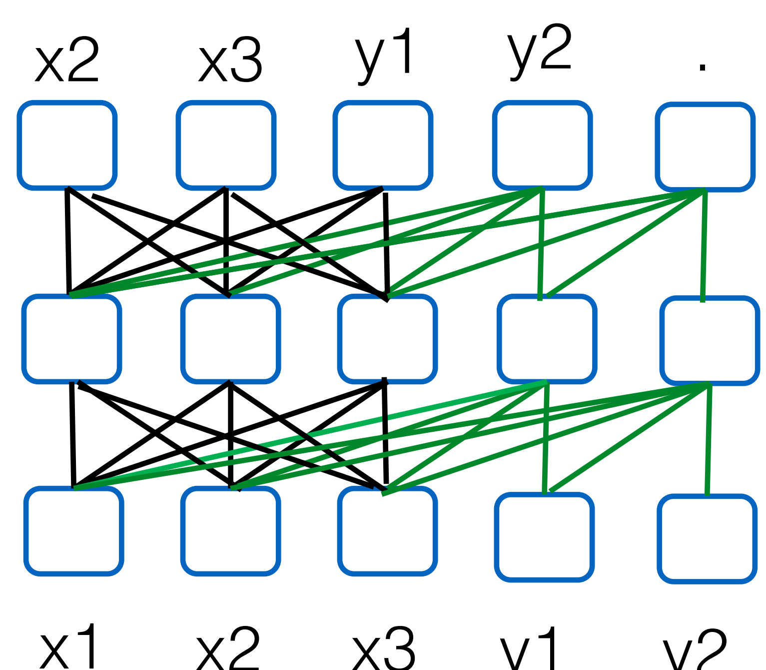
Seq2seq Pretraining and Beyond

Encoder-decoder



MASS/T5/BART

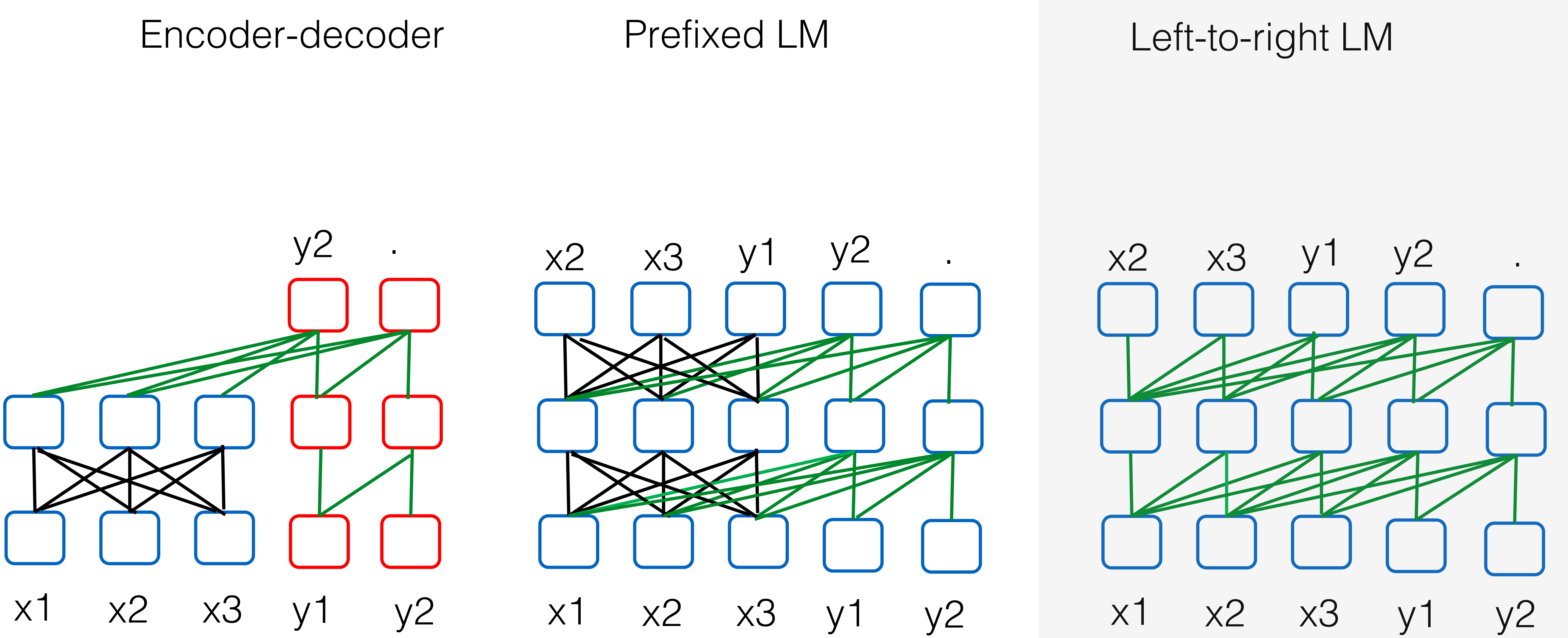
Prefixed LM



UNiLM/T5

?

Seq2seq Pretraining and Beyond



MASS/T5/BART

UNiLM/T5

GPT2/GPT3

GPT3

(Brown et al.)

- **Model:** like GPT2 (left-to-right, task prompt)
- **Objective:** left-to-right language model
- **Data:** Common Crawl + WebText + Books1,2 + Wikipedia

GPT3

(Brown et al.)

- Left-to-right language model is all you need:
    - with the help of “prompt”, many NLP tasks could be formulated as a language model problem.

Example

“Brevet Sans Garantie Du Gouvernement”

Translate into English

?

Example

“Brevet Sans Garantie Du Gouvernement, translated to English”

Prompt

GPT3

Patented without government warranty

Real Examples

GPT3 Authors a Python Function

Promp
t

PROMPT

```
def is_palendrome(s):
"""Check whether a string is a palindrome"""
```

Output

OUTPUT

```
s = s[::-1]

return s == s[::-1]
```

Image credit: https://www.buildgpt3.com/

Promp
t

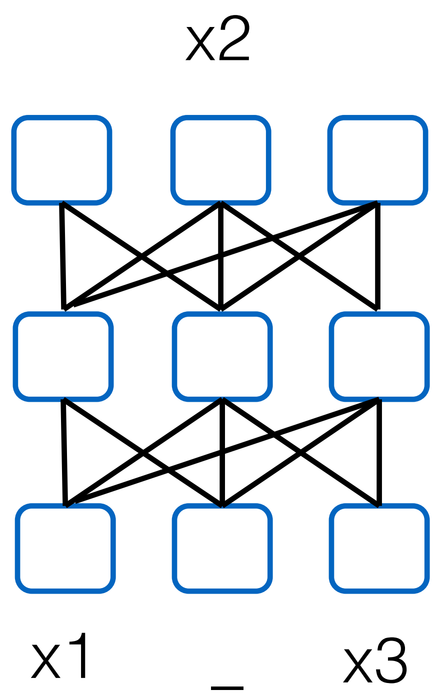
```
Q: What is human life expectancy in the United States?
A: Human life expectancy in the United States is 78 years.

Q: Who was president of the United States in 1955?
A: Dwight D. Eisenhower was president of the United States in 1955.

Q: What party did he belong to?
A: He belonged to the Republican Party.

Q: Who was president of the United States before George W. Bush?
A: Bill Clinton was president of the United States before George W. Bush.

Q: Who won the World Series in 1995?
A: The Atlanta Braves won the World Series in 1995.
```

Test

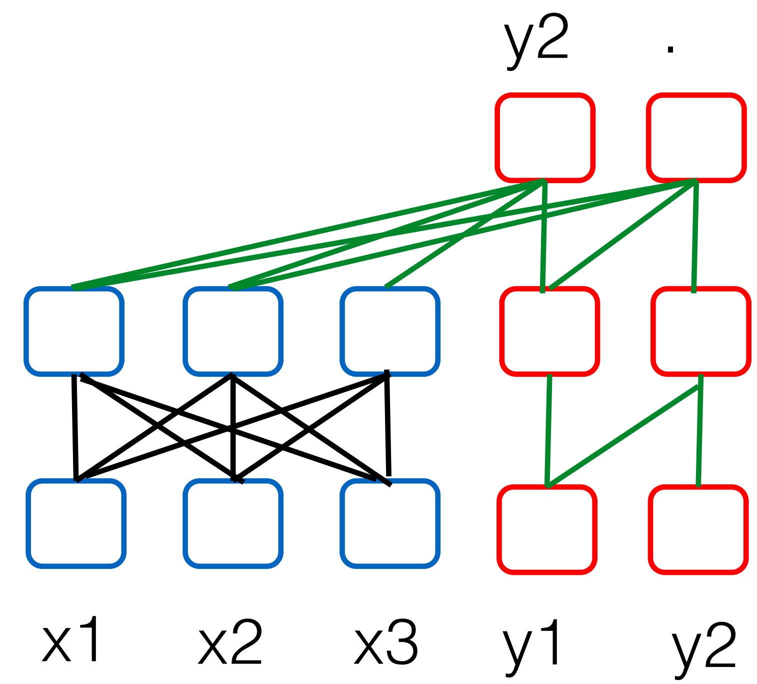
```
Q: What is your favorite animal?
A: My favorite animal is a dog.

Q: Why?
A: Because dogs are loyal and friendly.

Q: What are two reasons that a dog might be in a bad mood?
A: Two reasons that a dog might be in a bad mood are if it is hungry or if it is hot.

Q: How many eyes does a giraffe have?
A: A giraffe has two eyes.

Q: How many legs does a frog have?
A: A frog has four legs.

Q: Are there any animals with three legs?
A: No, there are no animals with three legs.

Q: Why don't animals have three legs?
A: Animals don't have three legs because they would fall over.
```

Example credit: https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html

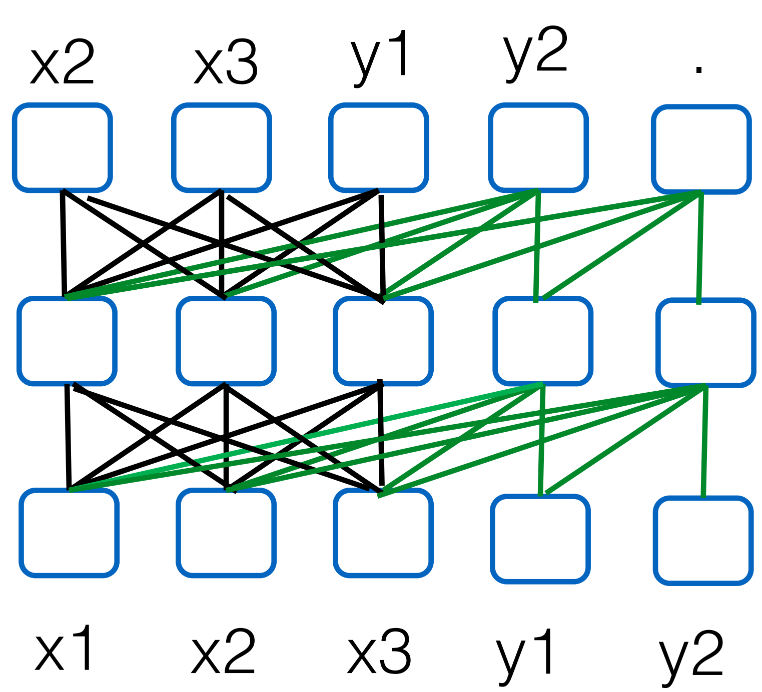
A Unified view for Pre-trained Models



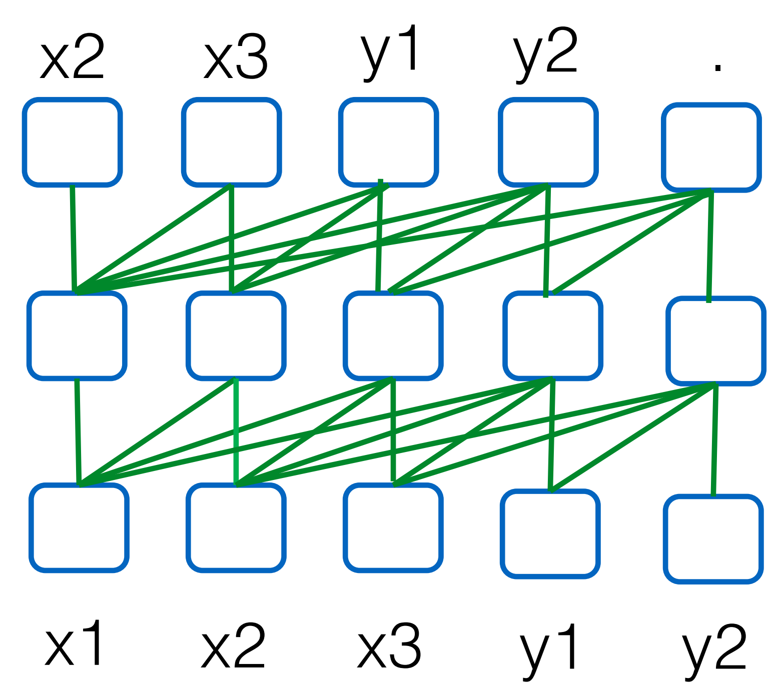
Masked LM



Encoder-decoder



Prefixed LM



Left-to-right LM

A Unified view for Pre-trained Models

	X			Y		
	Attention Mask	Corruption	Prediction objective	Attention Mask	Corruption	Prediction objective
Masked LM	Fully-connected	Mask token/span	Some words	-	-	-
Left-to-right LM	Autoregressive	none	All words	Autoregressive	none	All words
Prefix LM	Fully-connected	Mask token/span	Some words	Autoregressive	none	All words
Encoder-decoder	Fully-connected	Mask, drop, replace	No words	Autoregressive	none	All words

Practical Tricks

- Choose an appropriate one
    - Suitability: Data, loss of pre-train models <-> domain, characteristics of your tasks

Practical Tricks

- Choose an appropriate one
    - **Suitability**: Data, loss of pre-train models <-> domain, characteristics of your tasks

Empirically

- *Commonly, larger versions could do better.*
- *RoBERTa > BERT on many NLP tasks, but when designing a metric (BERTScore), BERT does better*
- *BART >> BERT on text generation tasks.*
- *GPT2 is suitable for unconditional text generation tasks.*
- *GPT3 does better on few-shot/zero-shot scenario.*

Practical Tricks

- Choose an appropriate one
  - Suitability: Data, loss of pre-train models <-> domain, characteristics of your tasks
  - Economy: different versions (base, large, huge) -> based on your computational resource
    - lighter version first
    - If the pre-trained model is too large to store for GPUs,
      - Think about distilled version
      - Think about data or model parallel

Practical Tricks

- When you're using them, think about
  - How are their data pre-processing methods?
    - Case sensitive/tokenize
  - Do you want to fine-tune or freeze them?
  - If fine-tuning, which types of fine-tuning methods you want to adopt?
    - Gradual unfreezing (Howard et al. 2018); Prefix-tuning (Li et al. 2021); P-tuning (Liu et al. 2021)
  - If your interested pre-trained model (say M) has already been fine-tuned by other relevant tasks (say M')?
    - If yes, you probably can use M' directly

Open Questions

- Data Contamination
    - whether test samples have already been seen during the pre-training stage

Open Questions

- Data Contamination
- Data Privacy
    - Pre-training data can be recovered from pre-training samples (Carlini et al. 2020)

Open Questions

- Data Contamination
- Data Privacy
- Downstream task specific pre-training
  - Data perspective

"pre-training on in-domain unlabeled data can improve performance on downstream tasks" (from T5)

Open Questions

- Data Contamination
- Data Privacy
- Downstream task specific pre-training
  - Data perspective
  - Loss function perspective
    - *PEGASUS (Zhang et al. 2019): Summarization-specific Pre-training Models*
    - *RefBERT (Varkel et al. 2020): Corefence-specific Pre-training Models*

Open Questions

- Data Contamination
- Data Privacy
- Downstream task specific pre-training
- How do we use it
  - Fine-tune or not?

Open Questions

- Data Contamination
- Data Privacy
- Downstream task specific pre-training
- How do we use it?
- Is it true that "large pre-trained model is all we need"?