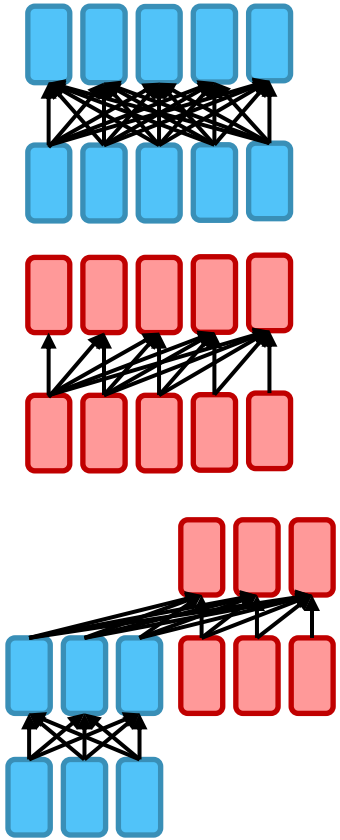
Applied Deep Learning

Model Pre-Training

April 25th, 2022 http://adl.miulab.tw

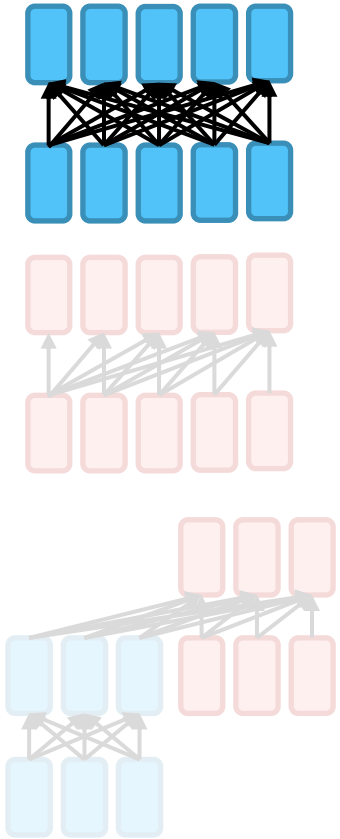
National Taiwan University
國立臺灣大學

Three Types of Model Pre-Training



- Encoder
  - Bidirectional context
  - Examples: BERT and its variants
- Decoder
  - Language modeling; better for generation
  - Example: GPT-2, GPT-3, LaMDA
- Encoder-Decoder
  - Sequence-to-sequence model
  - Examples: Transformer, BART, T5

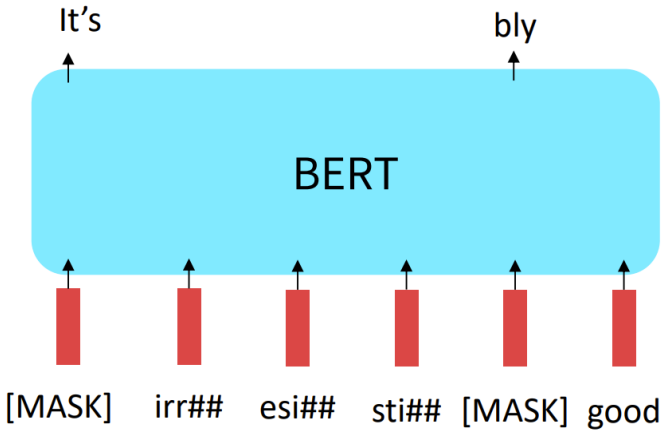
Three Types of Model Pre-Training

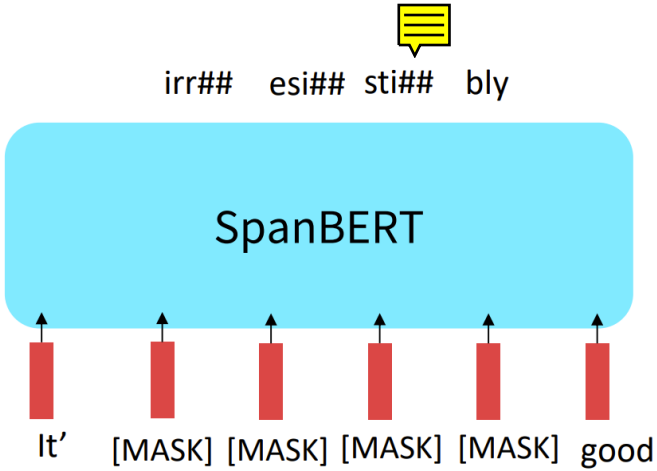


- Encoder
  - Bidirectional context
  - Examples: BERT and its variants
- Decoder
  - Language modeling; better for generation
  - Example: GPT-2, GPT-3, LaMDA
- Encoder-Decoder
  - Sequence-to-sequence model
  - Examples: Transformer, BART, T5

BERT Variants

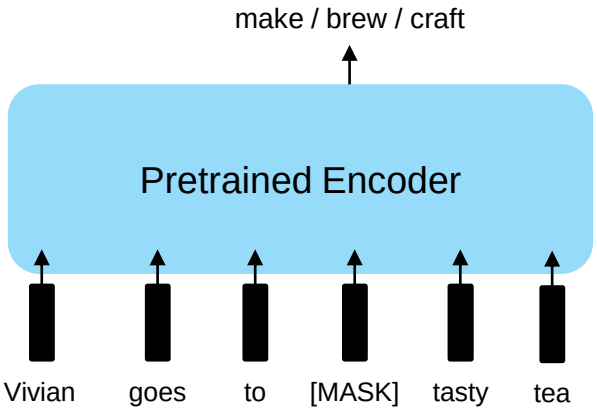
- Improvements to the BERT pretraining:
  - RoBERTa: mainly train BERT on *more data* and *longer*
  - SpanBERT: masking contiguous spans of words makes a harder, more useful pretraining task

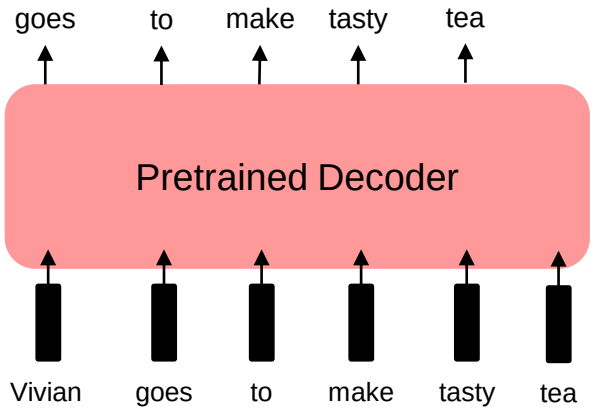




Need of Decoder

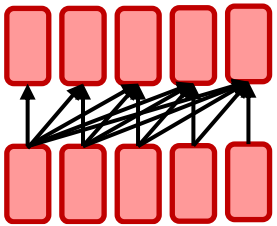
- Generating tasks
  - BERT and other pretrained encoders don’t naturally lead to *autoregressive (1-word-at-a-time)* generation methods





Three Types of Model Pre-Training

- Encoder
  - Bidirectional context
  - Examples: BERT and its variants

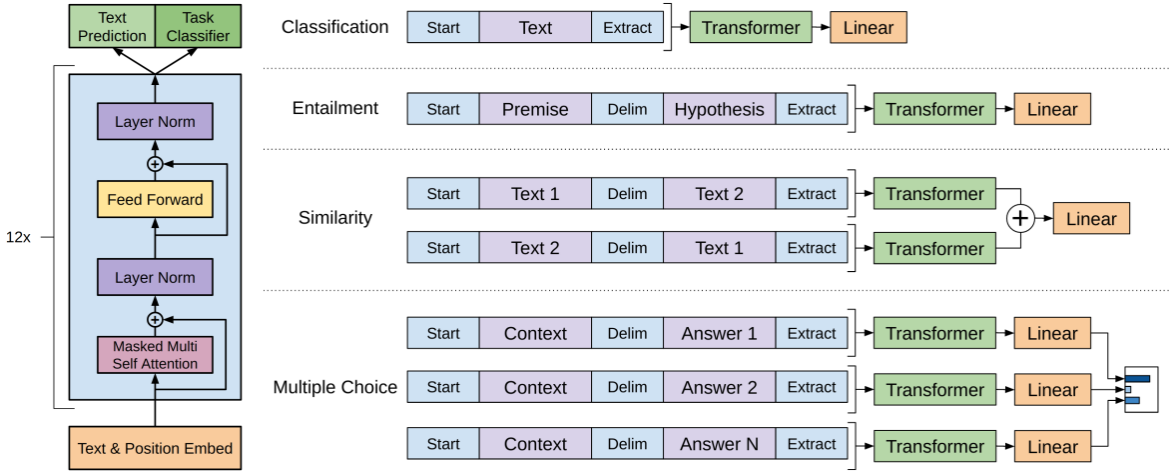


- **Decoder**
  - Language modeling; better for generation
  - Example: GPT-2, GPT-3, LaMDA

- Encoder-Decoder
  - Sequence-to-sequence model
  - Examples: Transformer, BART, T5

GPT: Generative Pretrained Transformer

- Transformer decoder
  - Pre-trained on BooksCorpus (~7000 books)
    - Transformer decoder with 12 layers
    - 768-dim hidden states, 3072-dim feed-forward hidden layers
    - Byte-pair encoding with 40,000 merges
  - Supervised fine-tuning for the target tasks



GPT-2

- Transformer decoder
  - Pre-trained on *more data*
  - Good for NLG

Context (WebText test)	
Outfit 1: Typical This pairing was the first outfit I thought of when I bought the shoes. It's like a summer version of this Jake Grantham outfit; in fact, my shoes are close to the colors of his Nike Racers! Instead of a heavy Harris Tweed jacket and denim shirt, I'm wearing a cotton DB jacket and and a linen shirt. Both fabrics (in these colors) are an absolute must for summer, as they go with both dark and and light pants! As you can see, they pair wonderfully with the dark jeans and shoes. It's a pseudo menswear/prep outfit. Overall, this is a very casual outfit which is why I paired my sneakers with it. I'm not about wearing a full wool suit with sneakers (as GQ shows a lot) but I'm definitely open to keeping things casual, like this cotton DB. Casual fabrics are key to pulling off your sneakers in a dressed down menswear outfit. I'd even suggest to wear these sneakers with a khaki chino suit or a white linen suit. Just be sure to ditch the tie or wear a tee or polo; wearing a tie with sneakers is a bit too much	
Small model completion	**GPT-2 completion**
. Outfit 2: High End This outfit was just in the eye of the storm. It's a hybrid of what ASICS does with streetwear (like J. Crew) and what J. Crew does with casual pieces that are a bit more traditional. Again, as I mentioned before, it's a low budget item and they can be very expensive. These will definitely be an issue if you're trying to work as a casual sneakermaker, and that's OK but don't hesitate if you want to try the lower price versions of their outfits! This version features a very low cut-cut neckline (which is not a bad thing), slightly longer legs, and a pair of long legs to keep things nice. My preferred color of color these items are blue-gray (which is my first colors, as they are a light white), or black (which is my favorite of my designs). As always, my pick of colors from this store have nothing to do with that other than the design. For a high end, I would recommend this to everyone! Outfit 3: High Light This was a "high light" look while I had a chance to get some color. I think the color is based on	for me! Outfit 2: Casual Jacket I know that some people might be opposed to wearing sneakers in a turtleneck, but I wanted to be true to myself, so I went with a slim fitting turtleneck in a color more similar to my favorite color of the day. Overall, it's a classic turtleneck, with nothing too flashy. I did swap my shoes for these ASICS Tiger Killshots and I wish I hadn't did that. The shoe itself is very comfortable, however, I found that wearing them in these sneakers made my feet look a bit larger. I'd wear them with denim, chinos, whatever, but would never recommend wearing them alone. There are just too many variables involved. Outfit 3: Dress This is the final outfit I wore the ASICS Tiger Killshots. This version of the ASICS is a much more modern and streamlined version. The shoe has the most modern and minimal design that I can think of! With a lot of black, these sneakers look very cool. They'll definitely be very comfortable for summer. To me, wearing them with a suit is the easiest way to have a good pair of sneaker shoes. That's pretty much it! If

More Powerful Pre-Trained Model – GPT-3

Pre-Training & Fine-Tuning

unannotated data

pre-train

Model

task-specific annotated data

fine-tune

Model

fine-tune

Model

Pre-Training & In-Context Learning

pre-train

Model

no learning

Model

no learning

Model

GPT-3 “In-Context” Learning

第一部份： 詞彙和結構
本部份共 15 題，每題含一個空格。請就試題冊上 A、B、C、D 四個選項中選出最適合題意的字或詞，標示在答案紙上。

題型說明

例：

It's eight o'clock now. Sue _____ in her bedroom.
A. study
B. studies
C. studied
D. is studying

正確答案為 D，請在答案紙上塗黑作答。

少數範例

GPT-3 “In-Context” Learning

- Zero-Shot

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

- One-Shot

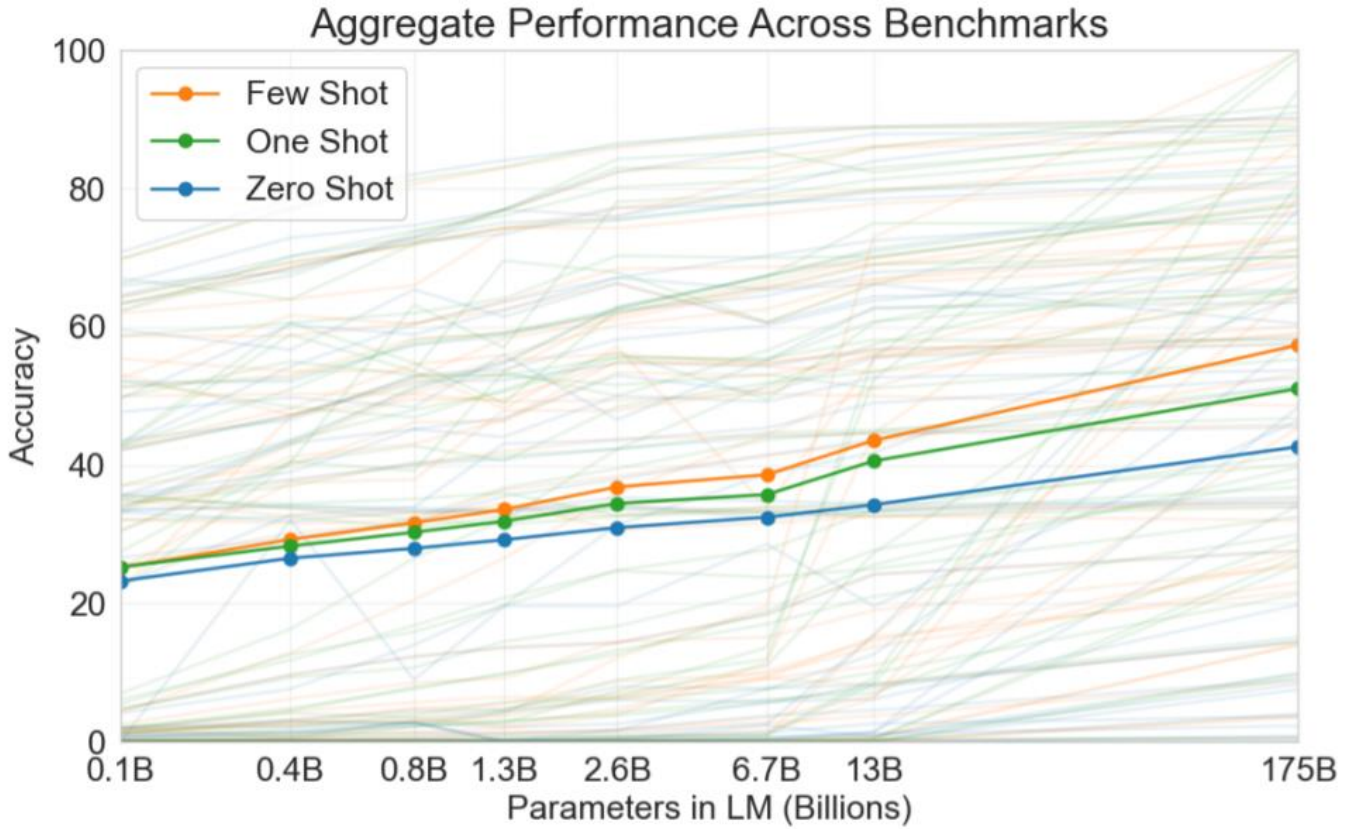
```
1  Translate English to French:      ← task description
2  sea otter => loutre de mer        ← example
3  cheese =>                         ← prompt
```

- Few-Shot

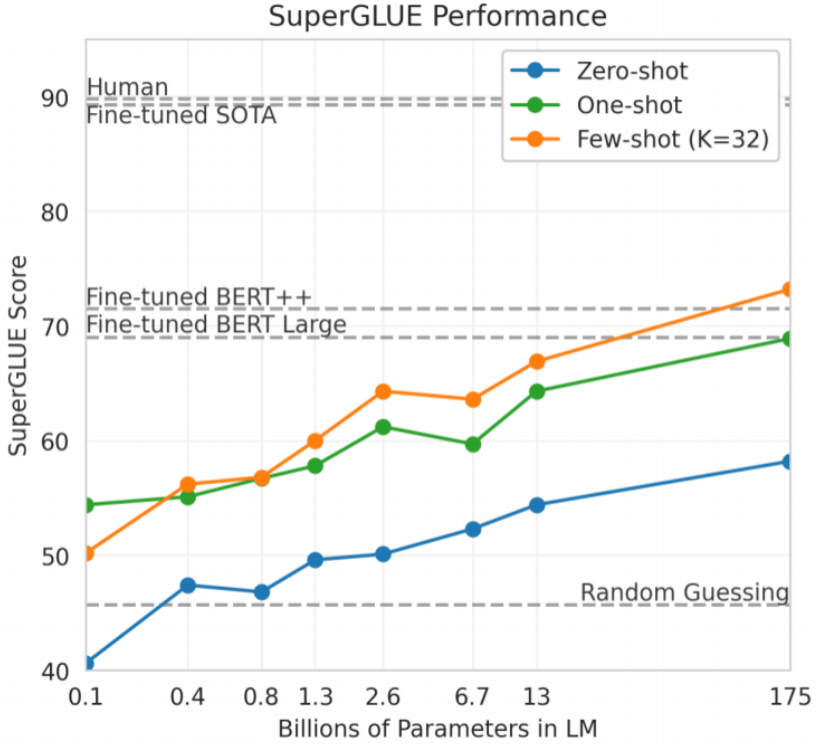
```
1  Translate English to French:      ← task description
2  sea otter => loutre de mer        ← examples
3  peppermint => menthe poivrée      ←
4  plush girafe => girafe peluche    ←
5  cheese =>                         ← prompt
```

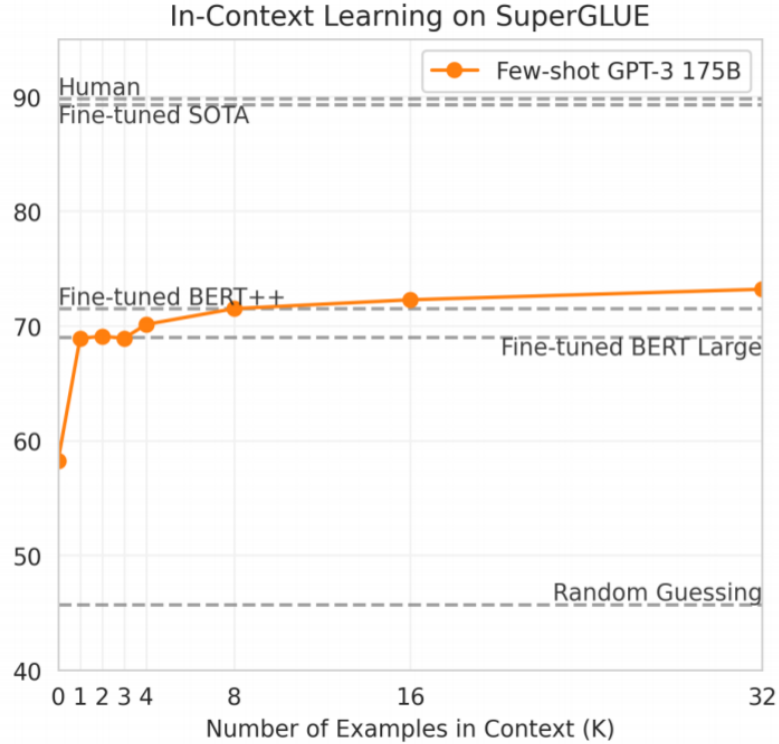
- Traditional Fine-Tuning

Benchmark 42 NLU Tasks



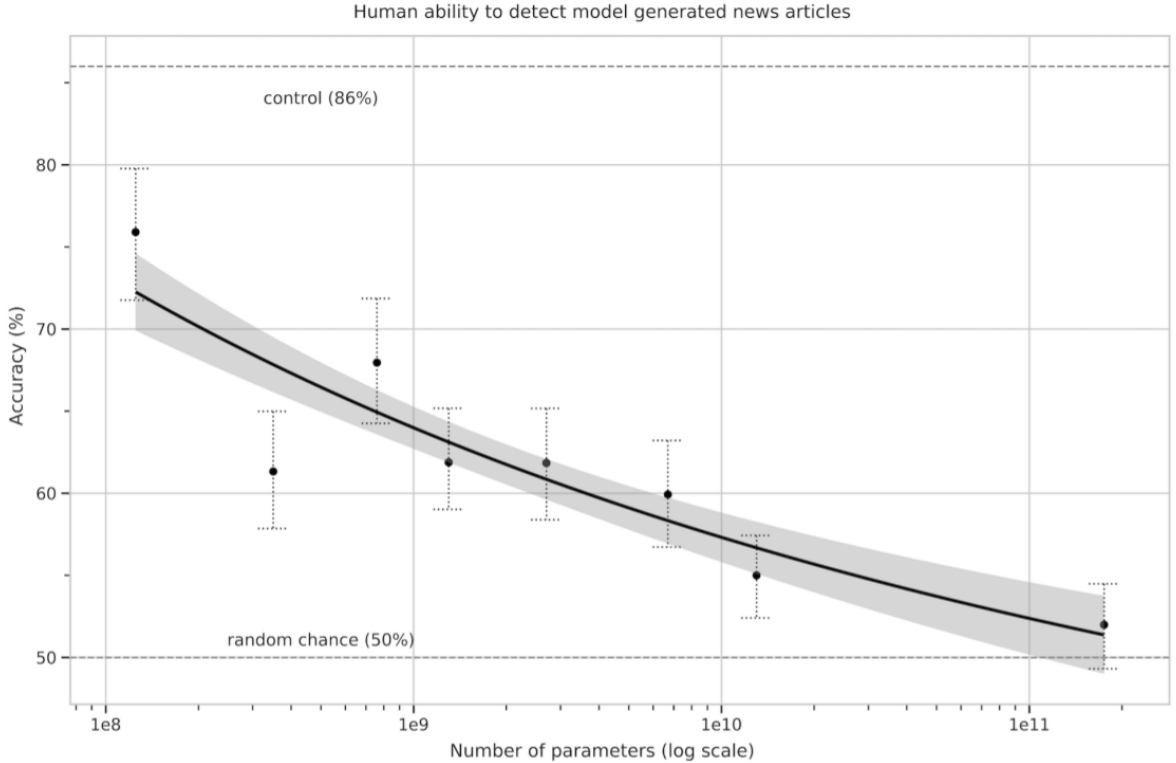
NLU Performance in SuperGLUE





NLG Performance

- Human identify if the article is generated



NLG Performance

- Using a new word in a sentence (few-shot)

A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:
We were traveling in Africa and we saw these very cute whatpus.

To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:
One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.
A "yalubalu" is a type of vegetable that looks like a big pumpkin. An example of a sentence that uses the word yalubalu is:
I was on a trip to Africa and I tried this yalubalu vegetable that was grown in a garden there. It was delicious.

A "Burringo" is a car with very fast acceleration. An example of a sentence that uses the word Burringo is:
In our garage we have a Burringo that my father drives to work every day.

A "Gigamuru" is a type of Japanese musical instrument. An example of a sentence that uses the word Gigamuru is:
I have a Gigamuru that my uncle gave me as a gift. I love to play it at home.

GPT-J Demo

- GPT-3 was released by openAI, has 175 billion parameters and is not openly available.
- GPT-J is a 6 billion parameter model released by Eleuther AI. The goal of the group is to democratize huge language models, so they released GPT-J and it is currently publicly available.
  - Better in code generation tasks

LaMDA: Language Models for Dialog Applications

- Pre-training: multiple public dialogue data (1.56T words)
- Fine-tuning: **Quality** and **Safety** scores
  - Using one model for both *generation* and *discrimination* enables an efficient combined *generate-and-discriminate* procedure.

    “<context><sentinel><response><attribute-name><rating>”

    - “What’s up? RESPONSE not much. SENSIBLE 1”
    - “What’s up? RESPONSE not much. INTERESTING 0”
    - “What’s up? RESPONSE not much. UNSAFE 0”

LaMDA: Language Models for Dialog Applications

- Fine-tuning for external knowledge via a tool set (TS)
  - Calculator: “135+7721”→ “7856”
  - Translator: “hello in French” → “Bonjour”
  - IR system: “How old is Rafael Nadal?” → “Rafael Nadal / Age / 35”

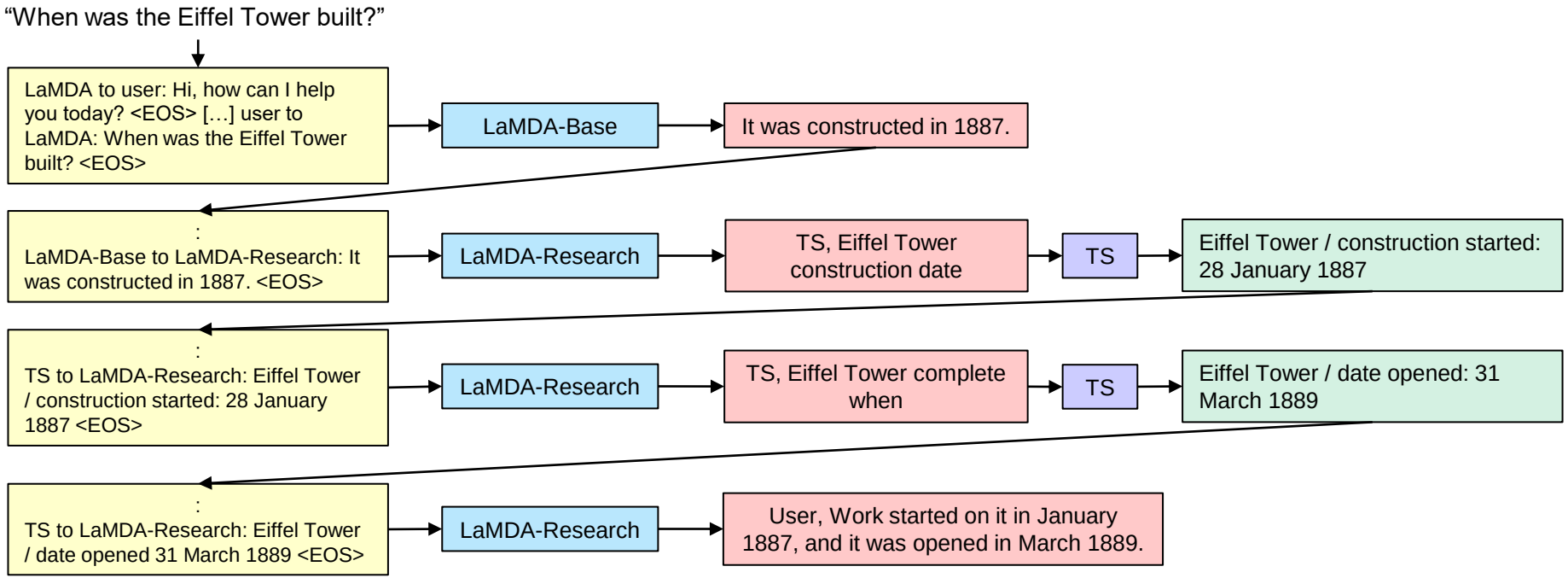
    context + base → “TS, Rafael Nadal’s age”
    - snippet: “He is 31 years old right now” + “Rafael Nadal / Age / 35”

    context + base + query + snippet → “User, He is 35 years old right now”

    context + base + query + snippet → “TS, Rafael Nadal’s favorite song”

- 40K dialog turns (generative data) are labeled ‘correct’ or ‘incorrect’ for the ranking task (discriminative data)

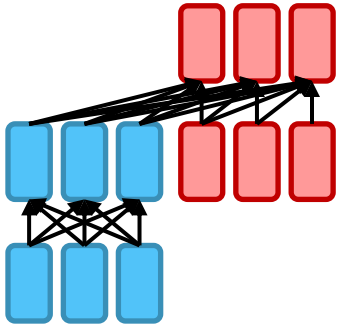
LaMDA Goundedness



Three Types of Model Pre-Training

- Encoder
  - Bidirectional context
  - Examples: BERT and its variants

- Decoder
  - Language modeling; better for generation
  - Example: GPT-2, GPT-3, LaMDA

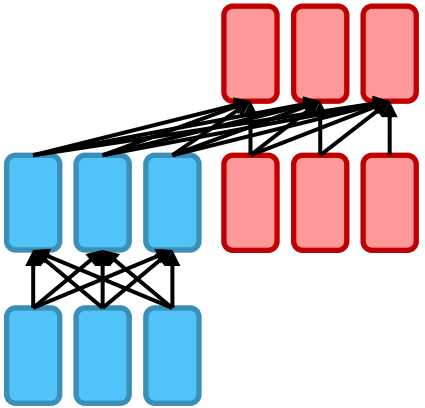


- Encoder-Decoder
  - Sequence-to-sequence model
  - Examples: Transformer, BART, T5

Encoder-Decoder Pre-Training

- The encoder portion benefits from bidirectional context; the decoder portion is used to train the whole model through language modeling.
- Pre-training objective: span corruption (denoising)
  - implemented in preprocessing
  - similar to language modeling at the decoder side

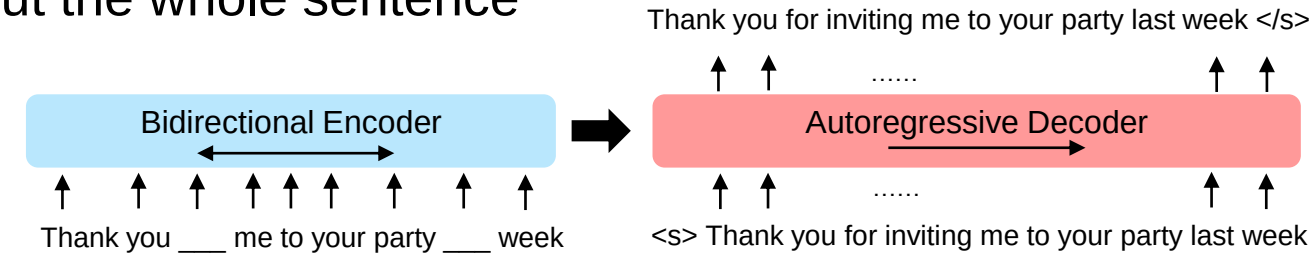
Thank you for ~~inviting me~~ to your party ~~last~~ week



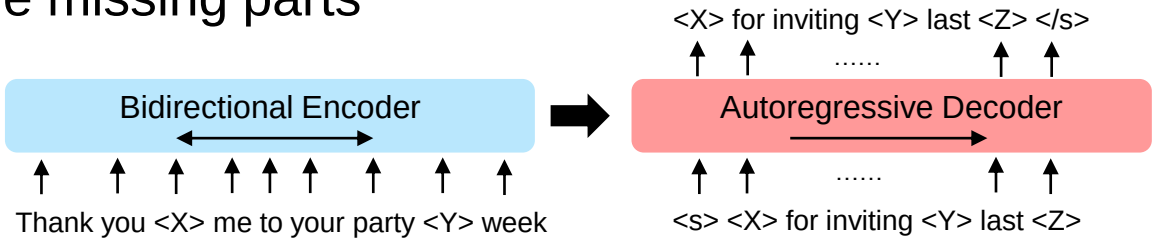
Denoising for Pre-Training

Thank you for ~~inviting me~~ to your party ~~last~~ week

- BART: output the whole sentence

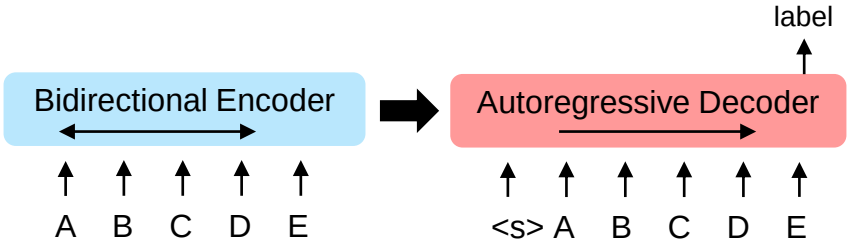


- T5: output the missing parts

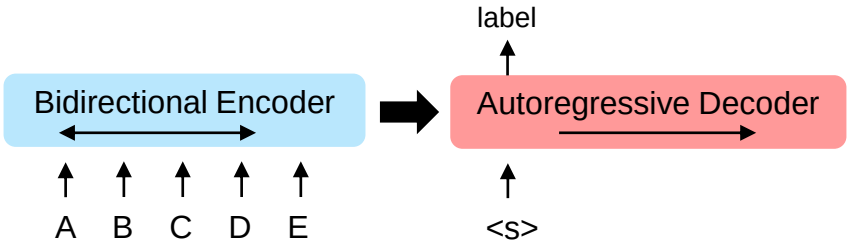


Fine-Tuning for Classification

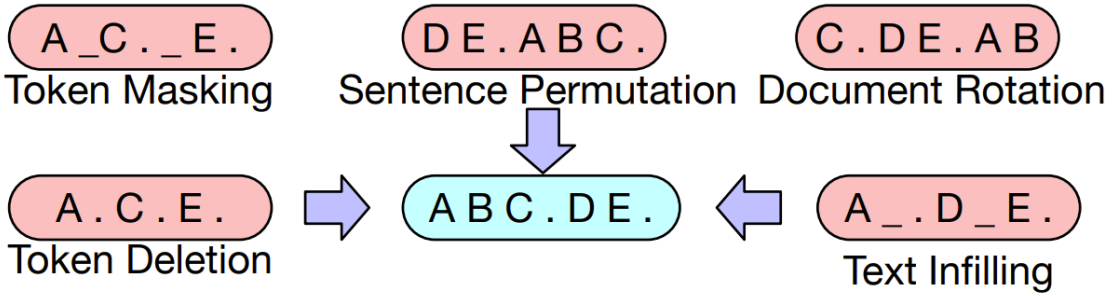
- BART: repeat input in decoder



- T5: treat it as a seq2seq task

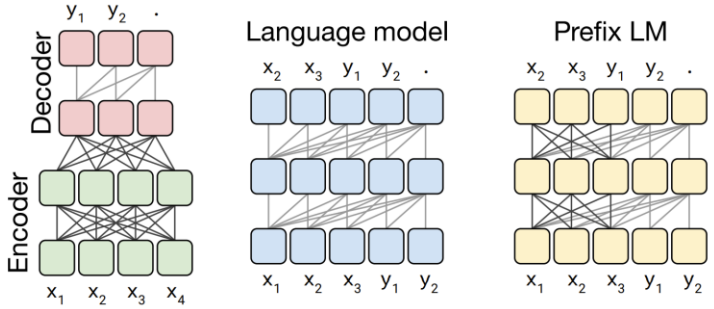


Diverse Noises in BART



Model	SQuAD 1.1 F1	MNLI Acc	ELI5 PPL	XSum PPL	ConvAI2 PPL	CNN/DM PPL
BERT Base (Devlin et al., 2019)	88.5	**84.3**	-	-	-	-
BART Base						
w/ Token Masking	90.4	84.1	25.05	7.08	11.73	6.10
w/ Token Deletion	90.4	84.1	24.61	6.90	11.46	5.87
w/ Text Infilling	**90.8**	84.0	24.26	**6.61**	**11.05**	5.83
w/ Document Rotation	77.2	75.3	53.69	17.14	19.87	10.59
w/ Sentence Shuffling	85.4	81.5	41.87	10.93	16.67	7.89
w/ Text Infilling + Sentence Shuffling	**90.8**	83.8	24.17	6.62	11.12	**5.41**

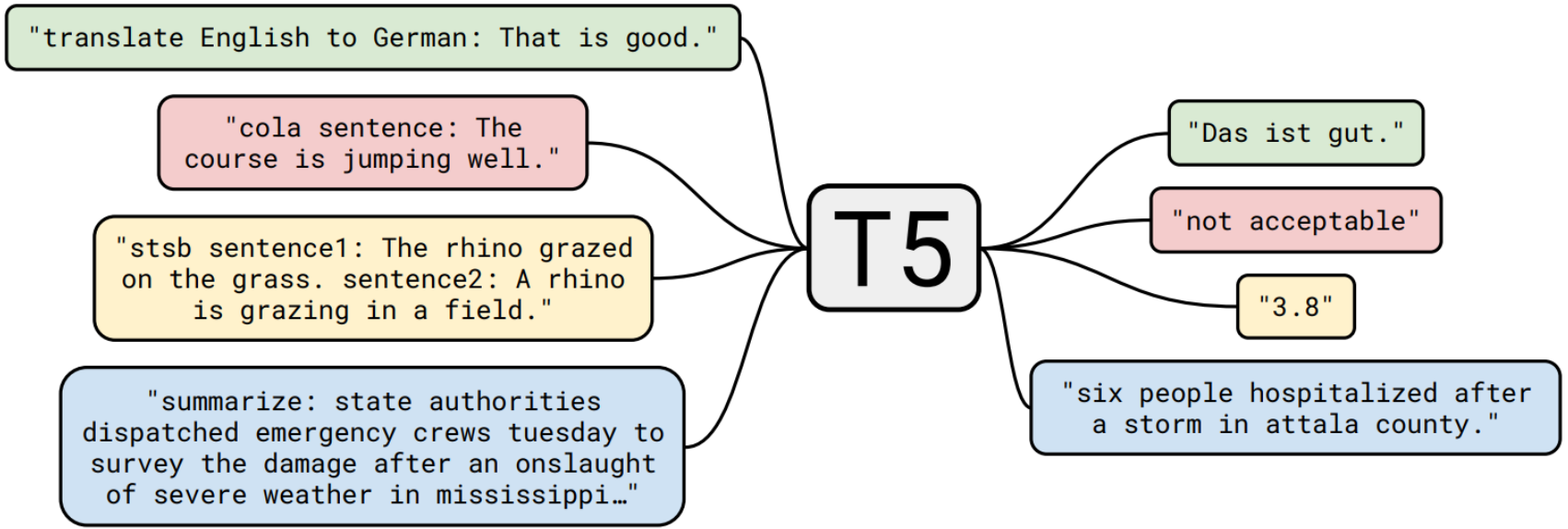
Effectiveness of Denoising in T5



Architecture	Objective	Params	Cost	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Encoder-decoder	Denoising	$2P$	M	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Enc-dec, shared	Denoising	P	M	82.81	18.78	**80.63**	**70.73**	26.72	39.03	**27.46**
Enc-dec, 6 layers	Denoising	P	$M/2$	80.88	18.97	77.59	68.42	26.38	38.40	26.95
Language model	Denoising	P	M	74.70	17.93	61.14	55.02	25.09	35.28	25.86
Prefix LM	Denoising	P	M	81.82	18.61	78.94	68.11	26.43	37.98	27.39
Encoder-decoder	LM	$2P$	M	79.56	18.59	76.02	64.29	26.27	39.17	26.86
Enc-dec, shared	LM	P	M	79.60	18.13	76.35	63.50	26.62	39.17	27.05
Enc-dec, 6 layers	LM	P	$M/2$	78.67	18.26	75.32	64.06	26.13	38.42	26.89
Language model	LM	P	M	73.78	17.54	53.81	56.51	25.23	34.31	25.38
Prefix LM	LM	P	M	79.68	17.84	76.87	64.86	26.28	37.51	26.76

T5: Text-to-Text Transfer Transformer

- **Multi-task pre-training**: learning multiple tasks via seq2seq



BART v.s. T5

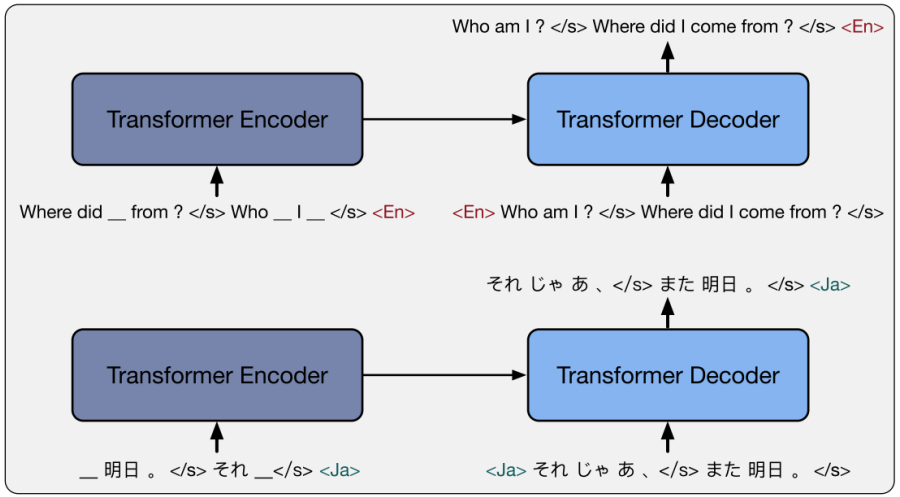
- Differences
  - Training data size: BART > T5 (about 2x)
  - Model size:
    - BART-large: 12 encoder, 12 decoder, 1024 hidden
    - T5-base: 12encoder, 12decoder, 768 hidden, 220M parameters (2x BERT-base)
    - T5-large: 24encoder, 24decoder, 1024hidden, 770M parameters
  - Position encoding: learnable absolute position (BART) & relative position (T5)
- Understanding performance

	SQuAD	MNLI	SST	QQP	QNLI	STS-B	RTE	MRPC	CoLA
BART	88.8 / 94.6	89.9 / 90.1	96.6	92.5	94.9	91.2	87.2	90.4	62.8
T5	86.7 / 93.8	89.9 / 89.6	96.3	89.9	94.8	89.9	87.0	89.9	61.2

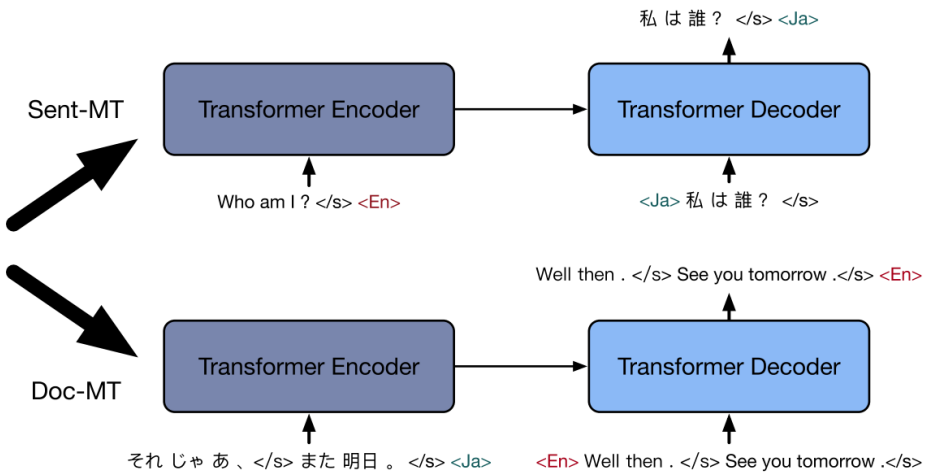
- Generation performance (summarization)

CNN/DailyMail	ROUGE-1	ROUGE-2	ROUGE-3
BART	45.14	21.28	37.25
T5	42.50	20.68	39.75

mBART: Multilingual BART



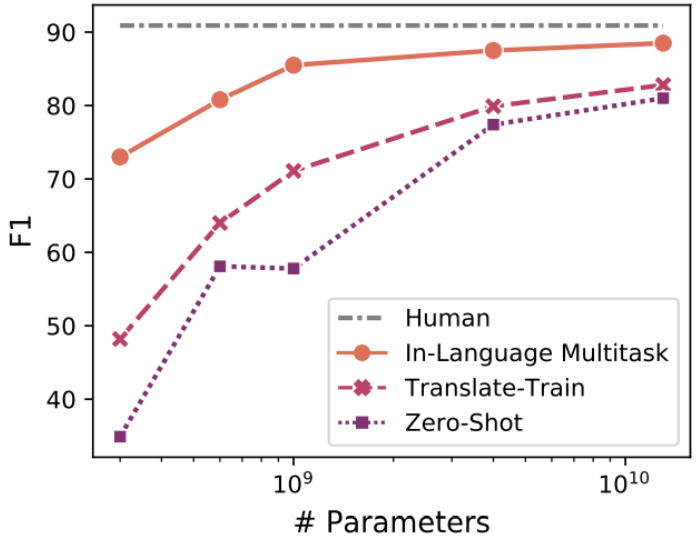
Multilingual Denoising Pre-Training (mBART)



Fine-tuning on Machine Translation

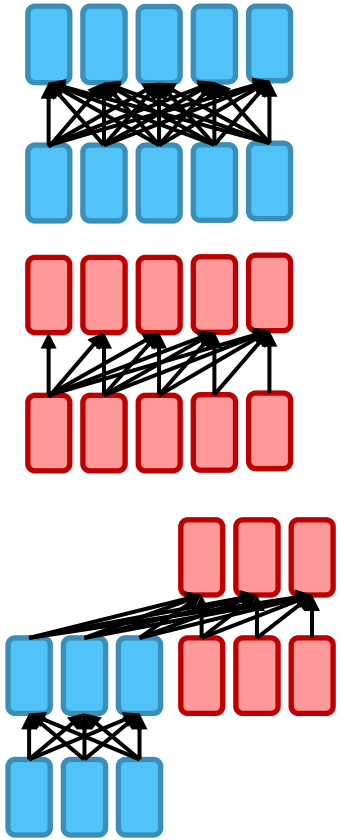
mT5: Multilingual T5

Model	Architecture
mBERT (Devlin, 2018)	Encoder-only
XLM (Conneau and Lample, 2019)	Encoder-only
XLM-R (Conneau et al., 2020)	Encoder-only
mBART (Lewis et al., 2020b)	Encoder-decoder
MARGE (Lewis et al., 2020a)	Encoder-decoder
mT5 (ours)	Encoder-decoder



Model	Sentence pair		Structured	Question answering		
	XNLI	PAWS-X	WikiAnn NER	XQuAD	MLQA	TyDiQA-GoldP
Metrics	Acc.	Acc.	F1	F1 / EM	F1 / EM	F1 / EM
Cross-lingual zero-shot transfer (models fine-tuned on English data only)						
mBERT	65.4	81.9	62.2	64.5 / 49.4	61.4 / 44.2	59.7 / 43.9
XLM	69.1	80.9	61.2	59.8 / 44.3	48.5 / 32.6	43.6 / 29.1
InfoXLM	81.4	-	-	- / -	73.6 / 55.2	- / -
X-STILTs	80.4	87.7	64.7	77.2 / 61.3	72.3 / 53.5	76.0 / 59.5
XLM-R	79.2	86.4	65.4	76.6 / 60.8	71.6 / 53.2	65.1 / 45.0
VECO	79.9	88.7	65.7	77.3 / 61.8	71.7 / 53.2	67.6 / 49.1
RemBERT	80.8	87.5	**70.1**	79.6 / 64.0	73.1 / 55.0	77.0 / 63.0
mT5-Small	67.5	82.4	50.5	58.1 / 42.5	54.6 / 37.1	35.2 / 23.2
mT5-Base	75.4	86.4	55.7	67.0 / 49.0	64.6 / 45.0	57.2 / 41.2
mT5-Large	81.1	88.9	58.5	77.8 / 61.5	71.2 / 51.7	69.9 / 52.2
mT5-XL	82.9	89.6	65.5	79.5 / 63.6	73.5 / 54.5	75.9 / 59.4
mT5-XXL	**85.0**	**90.0**	69.2	**82.5 / 66.8**	**76.0 / 57.4**	**80.8 / 65.9**
Translate-train (models fine-tuned on English data plus translations in all target languages)						
XLM-R	82.6	90.4	-	80.2 / 65.9	72.8 / 54.3	66.5 / 47.7
FILTER + Self-Teaching	83.9	91.4	-	82.4 / 68.0	76.2 / 57.7	68.3 / 50.9
VECO	83.0	91.1	-	79.9 / 66.3	73.1 / 54.9	75.0 / 58.9
mT5-Small	64.7	79.9	-	64.3 / 49.5	56.6 / 38.8	48.2 / 34.0
mT5-Base	75.9	89.3	-	75.3 / 59.7	67.6 / 48.5	64.0 / 47.7
mT5-Large	81.8	91.2	-	81.2 / 65.9	73.9 / 55.2	71.1 / 54.9
mT5-XL	84.8	91.0	-	82.7 / 68.1	75.1 / 56.6	79.9 / 65.3
mT5-XXL	**87.8**	**91.5**	-	**85.2 / 71.3**	**76.9 / 58.3**	**82.8 / 68.8**
In-language multitask (models fine-tuned on gold data in all target languages)						
mBERT	-	-	89.1	-	-	77.6 / 68.0
mT5-Small	-	-	83.4	-	-	73.0 / 62.0
mT5-Base	-	-	85.4	-	-	80.8 / 70.0
mT5-Large	-	-	88.4	-	-	85.5 / 75.3
mT5-XL	-	-	90.9	-	-	87.5 / 78.1
mT5-XXL	-	-	**91.2**	-	-	**88.5 / 79.1**

Concluding Remarks



- Encoder
  - Bidirectional context
  - Examples: BERT and its variants
- Decoder
  - Language modeling; better for generation
  - Example: GPT-2, GPT-3, LaMDA
- Encoder-Decoder
  - Sequence-to-sequence model
  - Examples: Transformer, BART, T5