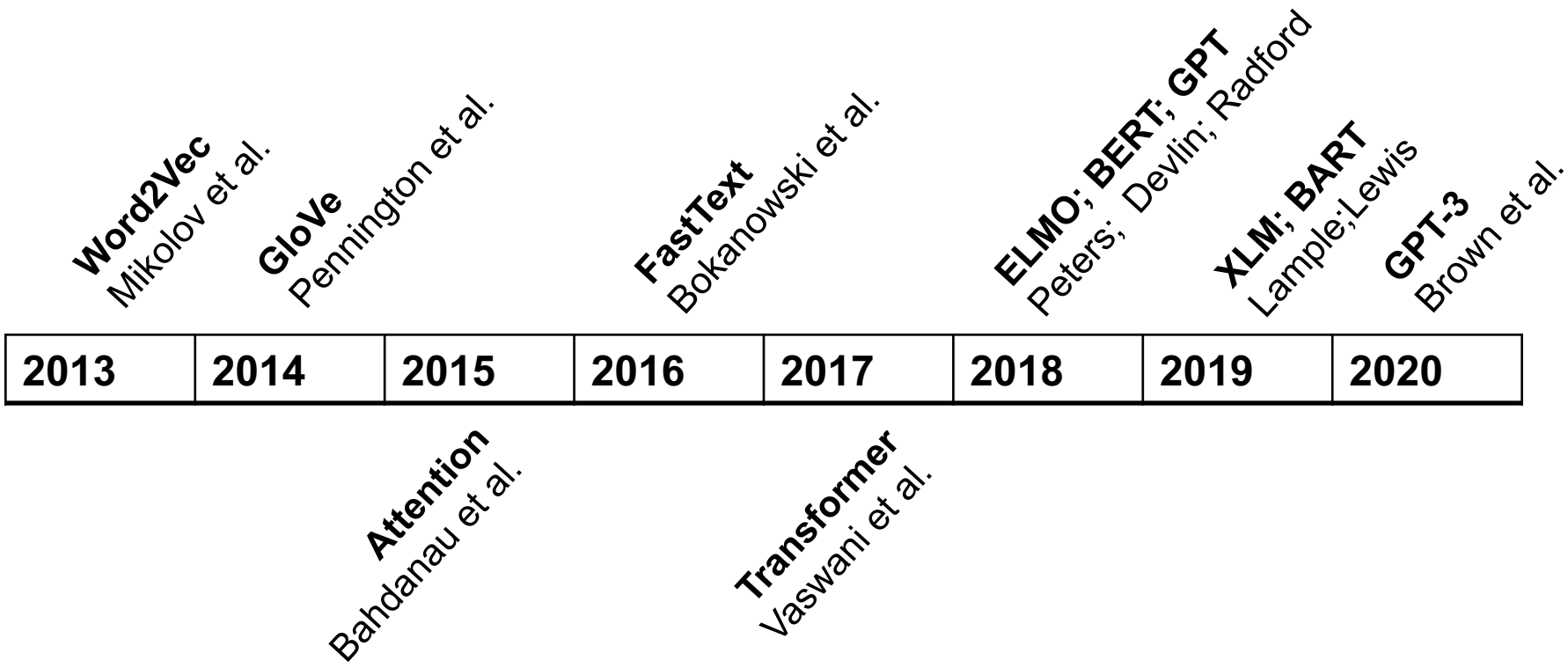
Advanced Transformers:

Pre-trained Language Modeling

ELMo, BERT, GPT and beyond

Marta Ruiz Costa-jussà, José Adrián Rodríguez Fonnollosa and Noé Casas

Timeline


Word2Vec
Mikolov et al.
GloVe
Pennington et al.
FastText
Bokanowski et al.
ELMO; BERT; GPT
Peters; Devlin; Radford
XLM; BART
Lample;Lewis
GPT-3
Brown et al.
2013
2014
2015
2016
2017
2018
2019
2020
Attention
Bahdanau et al.
Transformer
Vaswani et al.

Outline

- Pre-trained Language Modeling
  - Motivation: contextual word embeddings
  - Background: LM and Transformers
  - ELMO, BERT, GPT
- Latest Model GPT-3
- Evaluation

Contextual embeddings: intuition

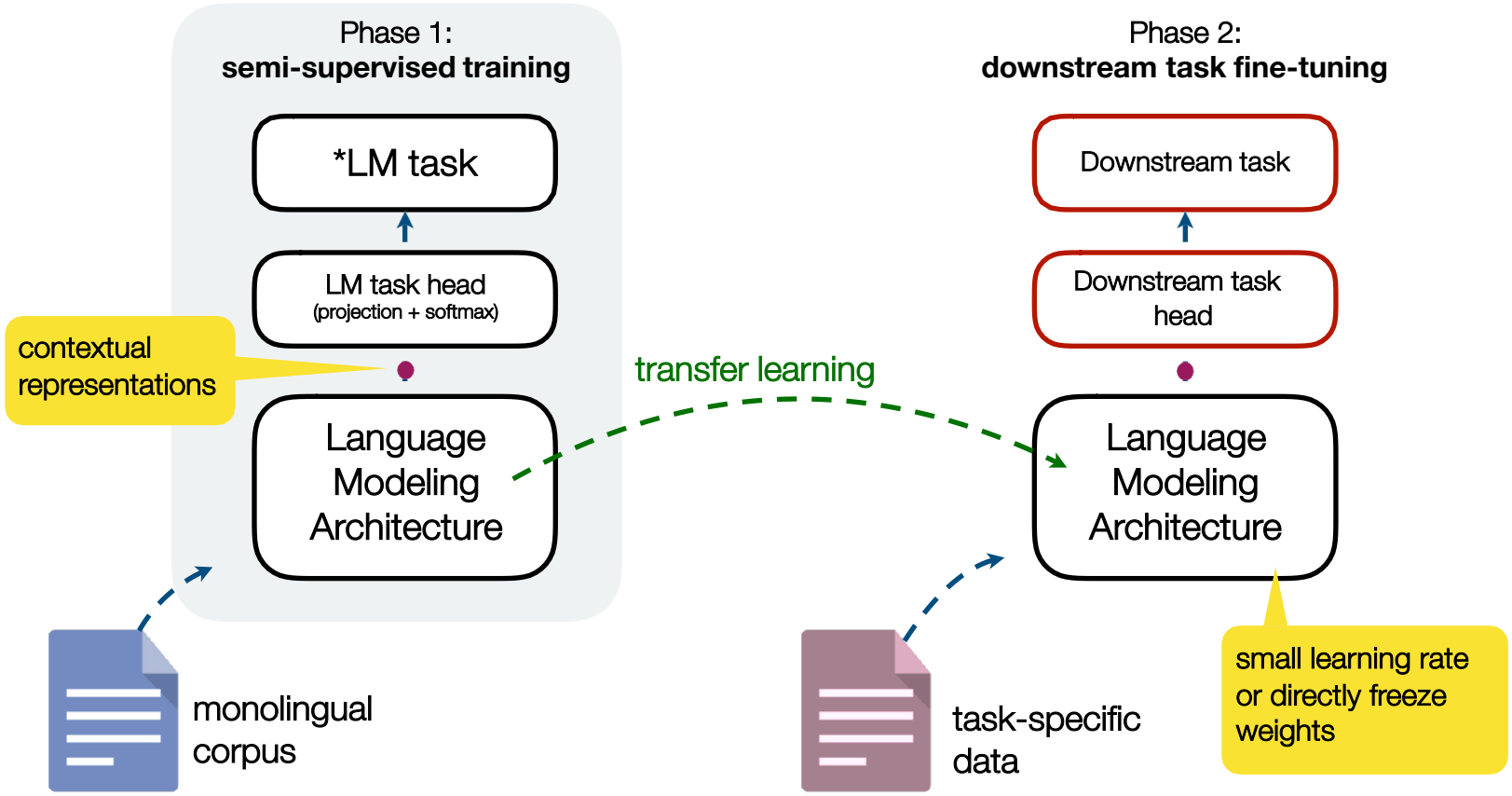
- Same word can have different meaning depending on the context. Example:
  - *Please,* ***type*** *everything in lowercase.*
  - *What* ***type*** *of flowers do you like most?*
- Classic word embeddings offer the same vector representation regardless of the context.
- Solution: create word representations that depend on the context.

Why Contextual Embeddings?

- Train model in one of multiple tasks that lead to word representations.
- Release pre-trained models.
- Use pre-trained models, options:
  A. Fine-tune model on final task.
  B. Directly encode token representations with model.

Explaining why graphically

downstream **tasks is what the field calls those supervised-learning tasks that utilize a pre-trained model or component**

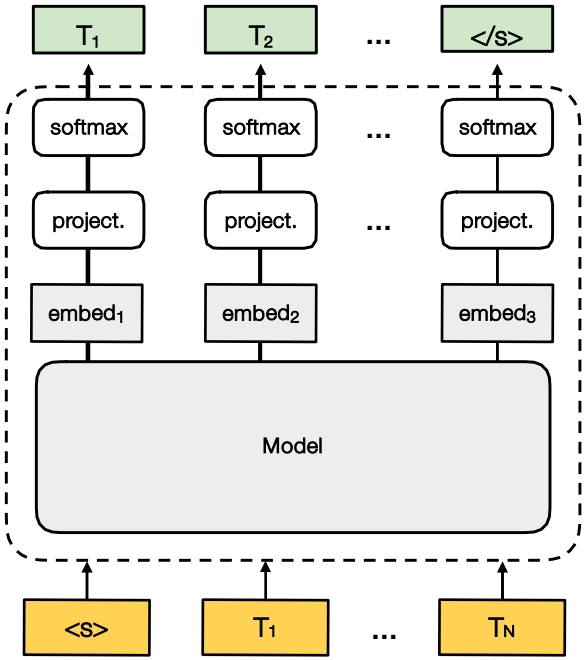


Background: Language Modeling

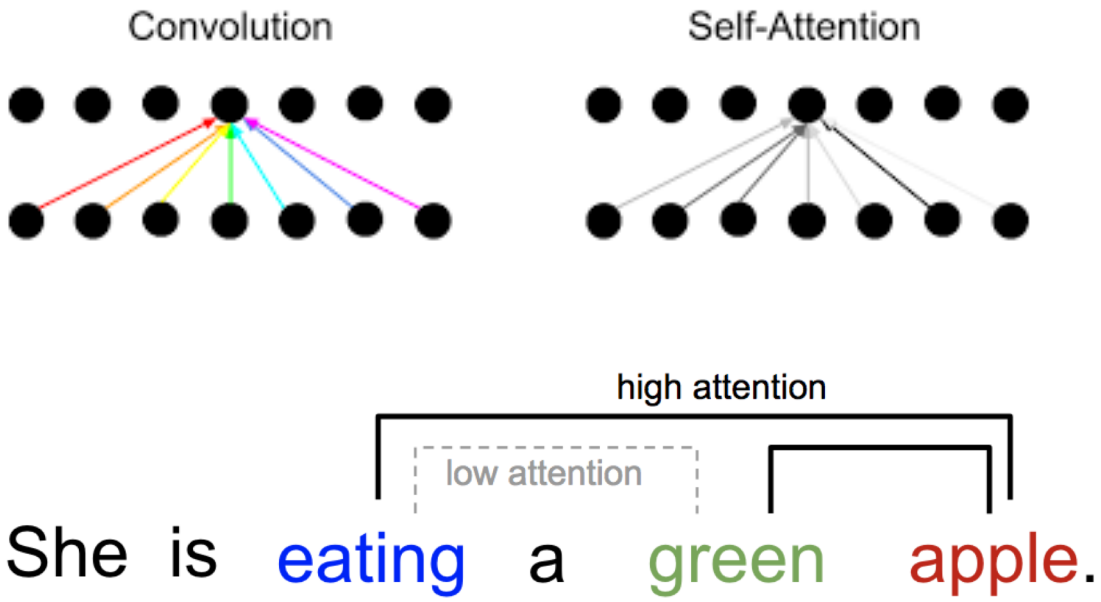
- Data: Monolingual Corpus
- Task: predict next token given previous tokens (causal):

$$P(T_i \mid T_1 \ldots T_{i-1})$$

- Usual models: LSTM, Transformer.



Non-recurrent attention-based language models



Transformer Language Model (Self-attention)

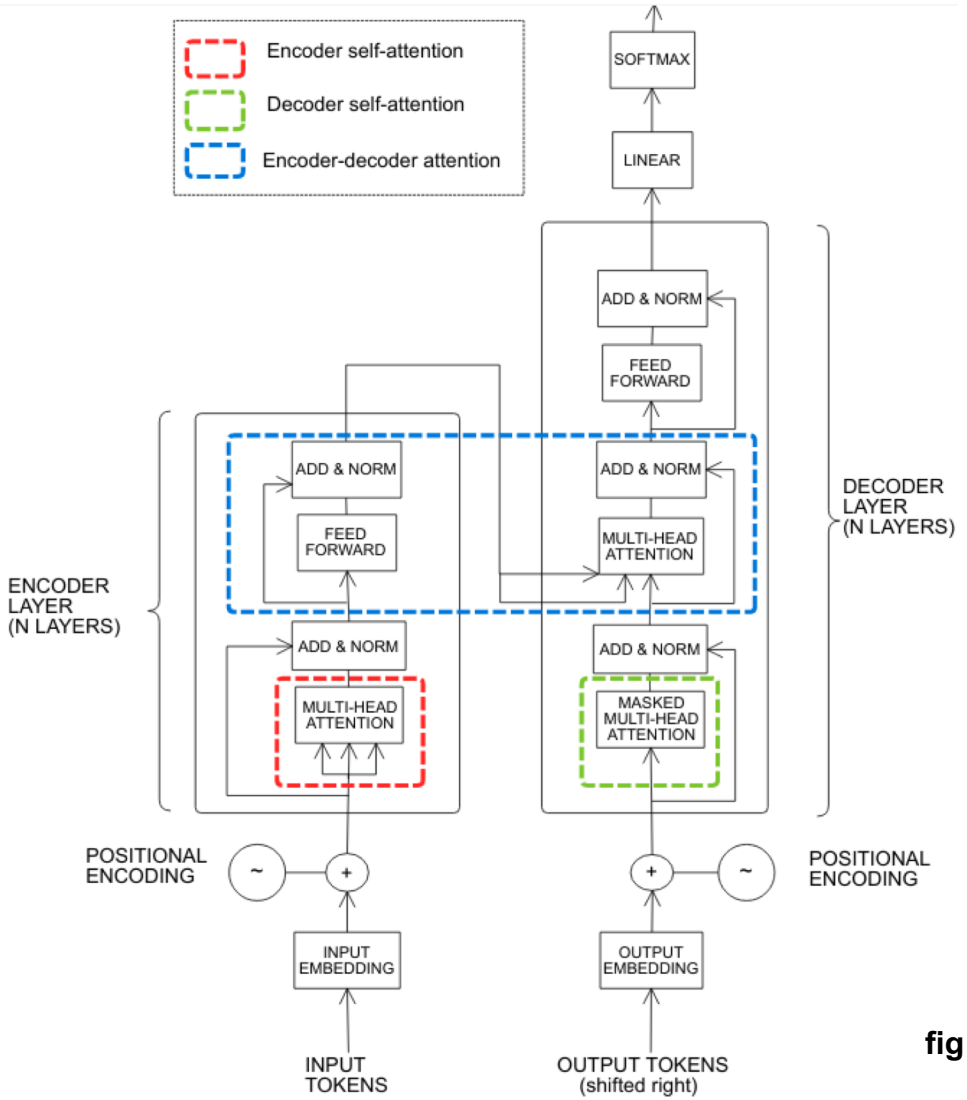
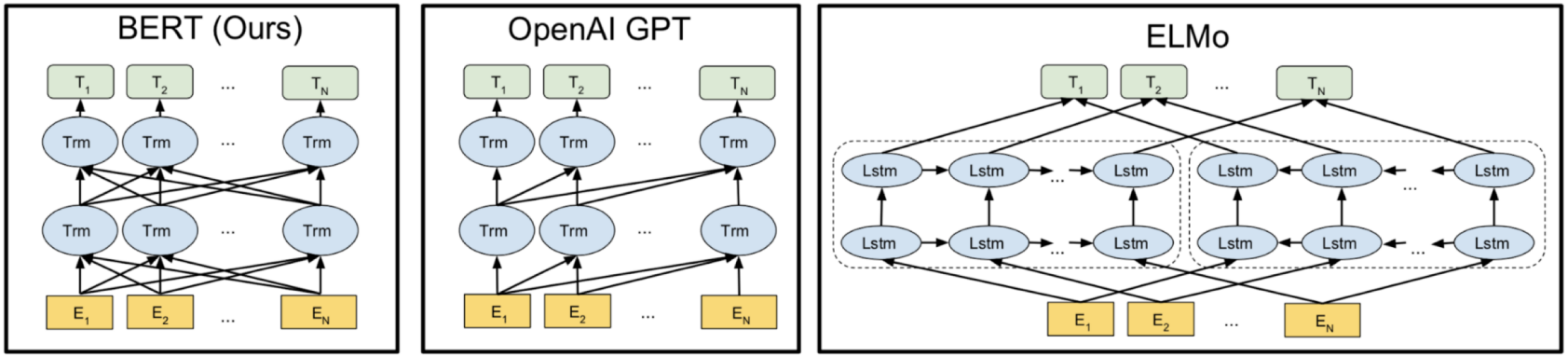


fig: Jordi armengol

Contextual word embeddings

BERT, GPT, ELMO



"Masked language models" can be used for any NLP task as "contextual word embeddings"

Articles

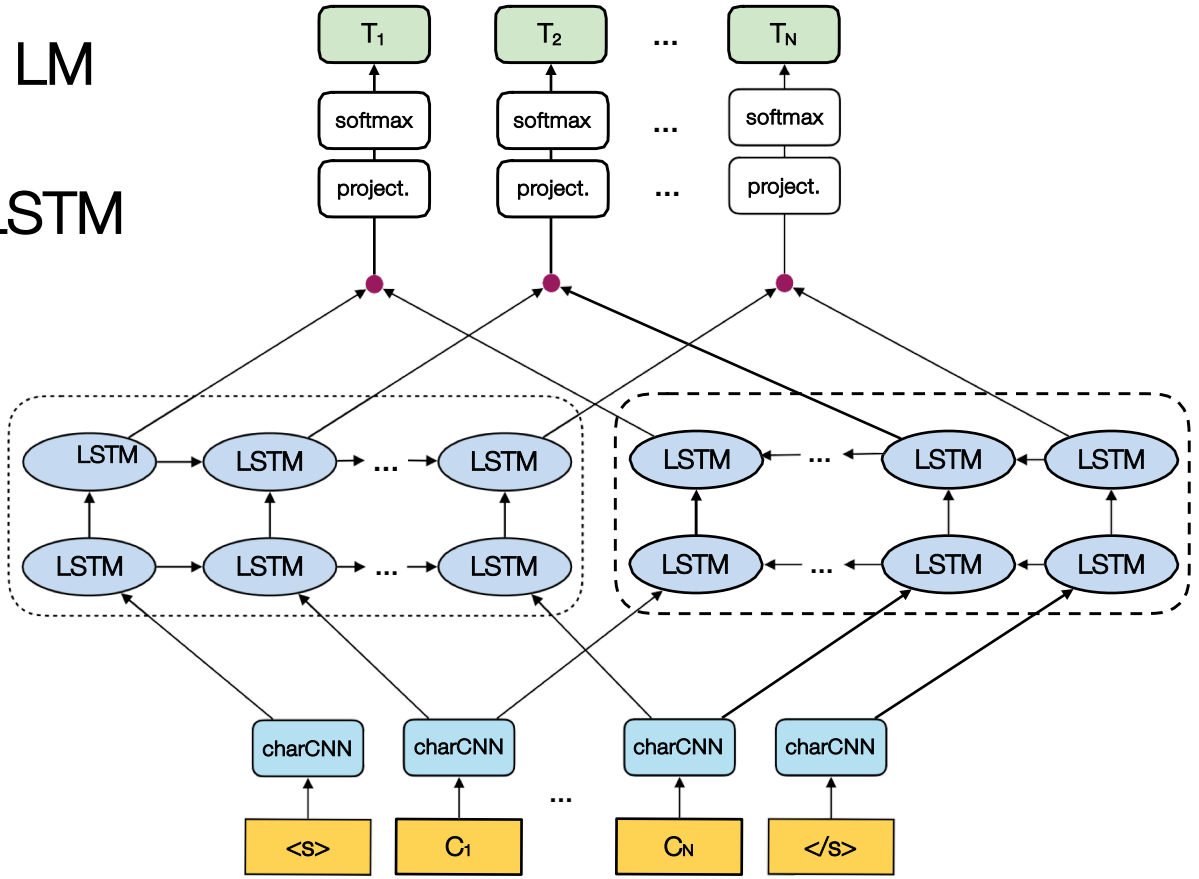
Model Alias	Org.	Article Reference
ELMo	AllenNLP	*Deep contextualized word representations* Peters et al.
OpenAI GPT	OpenAI	*Improving Language Understanding by Generative Pre-Training* Radford et al.
BERT	Google	*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* Devlin et al.

Differences

Alias	Model	Token	Tasks	Language
ELMo	LSTM	word	Bidirectional LM	English
OpenAI GPT	Transformer decoder	subword	Causal LM + Classification	English
BERT	Transformer encoder	subword	Masked LM + Next sentence prediction	Multilingual

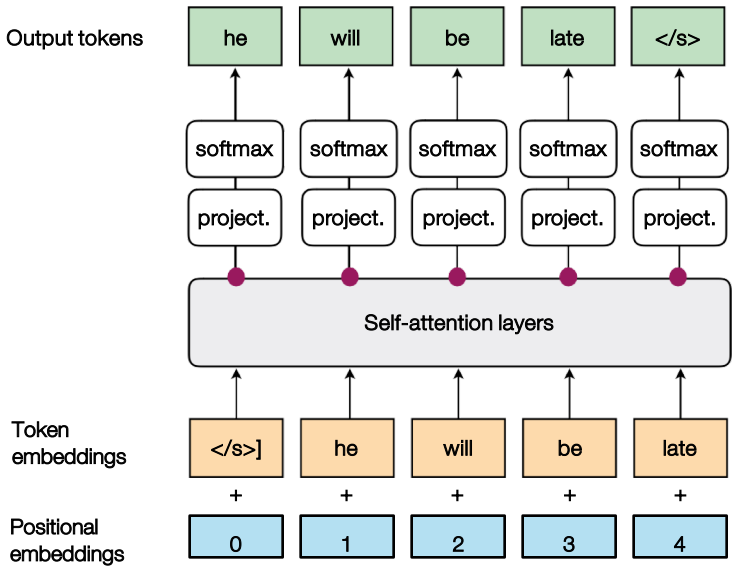
ELMO: contextual word representations

- **Task**: bidirectional LM
- **Model**: 2-layer biLSTM
- **Tokens**: words



OpenAI GPT: Generative Pre-Training

- **Task**: causal LM
- **Model**: self-attention layers
- **Tokens**: subwords
- Transformer Decoder



learning a generative language model using unlabeled data and then fine-tuning the model by providing examples of specific downstream tasks

More on GPT

- Learning objectives and concepts:
  - unsupervised language modelling (pre-training)

$$L_1(T) = \sum_i log\, P(t_i|t_{i-k}, \ldots, t_{i-1}; \theta) \qquad (i)$$

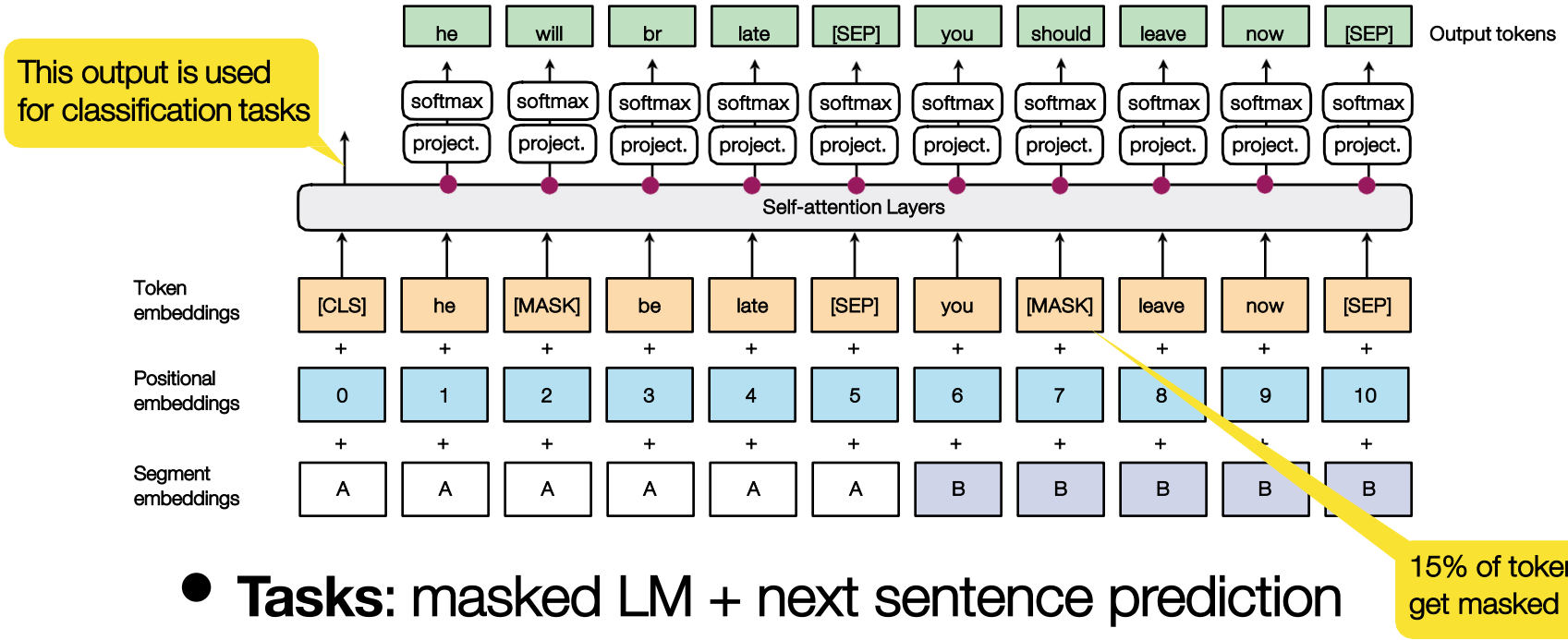
  - supervised fine-tuning and modified training objective

$$L_2(C) = \sum_{x,y} log\, P(y|x_1, \ldots, x_n) \qquad (ii)$$

$$L_3(C) = L_2(C) + \lambda L_1(C) \qquad (iii)$$

tasks like textual entailment, semantic similarity, question answering and commonsense reasoning,

BERT

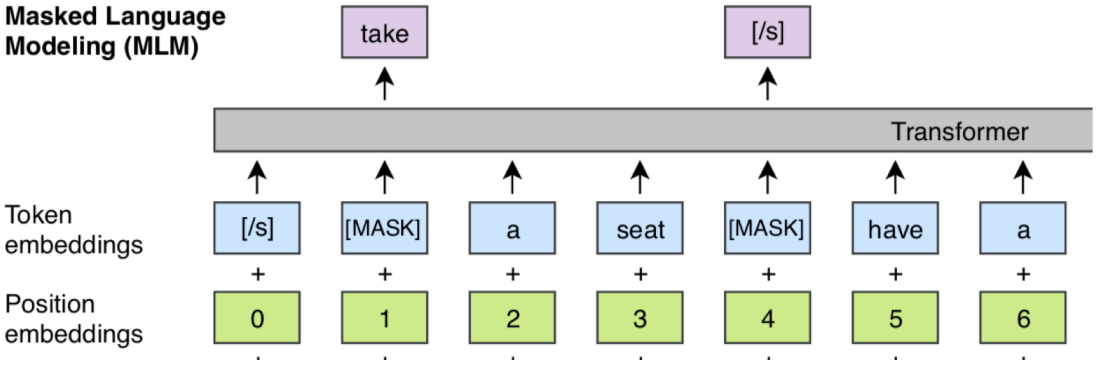


- **Tasks**: masked LM + next sentence prediction
- **Model**: self-attention layers
- **Tokens**: subwords

- Transformer Encoder

Masked language modeling task

BERT is based on “gap-fill” exercises (Transformer)



Mask out k% of the input words, and then predict the masked words

- They always use k = 15%

store

- Too little masking: Too expensive to train
- Too much masking: Not enough context

Sherlock Holmes is probably the most famous detective in **MASK**. Of course, he wasn't a real person. His **MASK** is based on a real man whose career had a great influence on Arthur Conan Doyle, the author of the detective stories.

https://arxiv.org/abs/1810.04805

Next sentence prediction task

- To learn relationships between sentences, predict whether Sentence B is actual sentence that proceeds Sentence A, or a random sentence

Sentence A = `The man went to the store.`
Sentence B = `He bought a gallon of milk.`
Label = `IsNextSentence`

Sentence A = `The man went to the store.`
Sentence B = `Penguins are flightless.`
Label = `NotNextSentence`

Contextual word embeddings. BERT

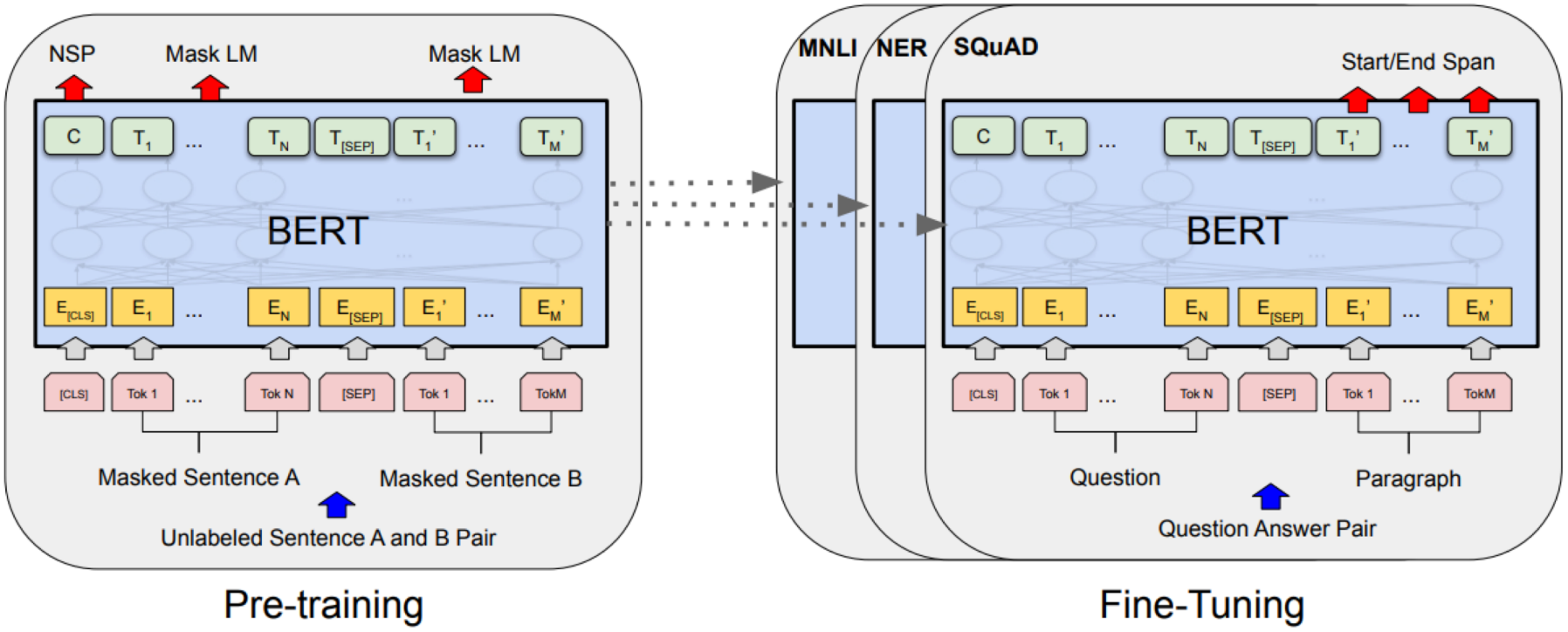
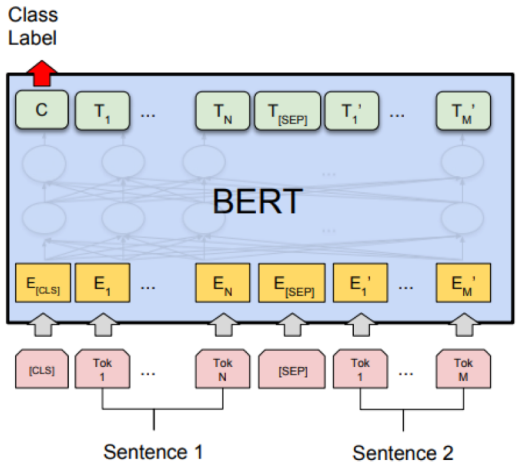


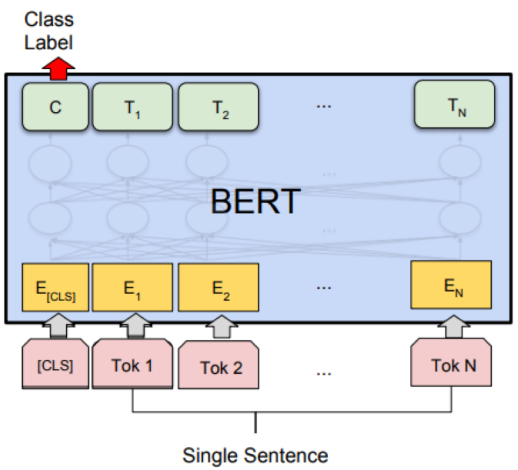
Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. `[CLS]` is a special symbol added in front of every input example, and `[SEP]` is a special separator token (e.g. separating questions/answers).

https://arxiv.org/abs/1810.04805

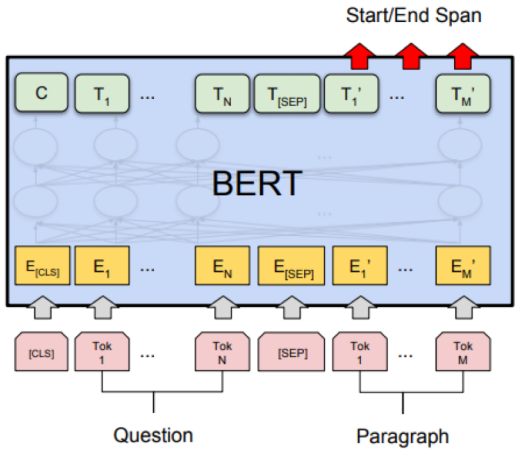
Contextual word embeddings. BERT



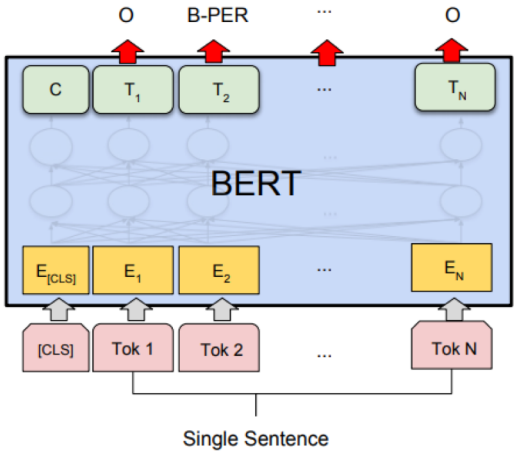
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



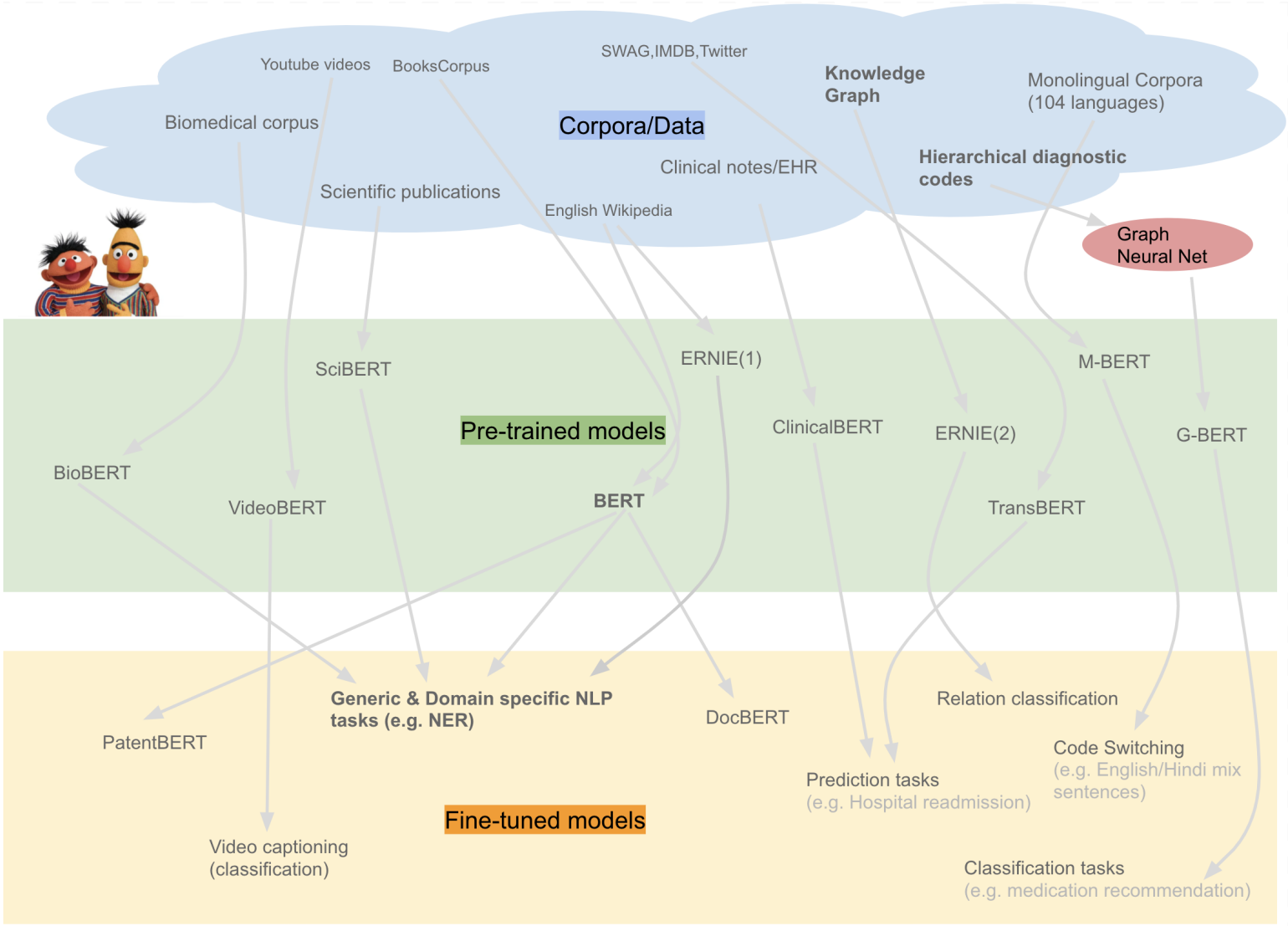
(c) Question Answering Tasks: SQuAD v1.1



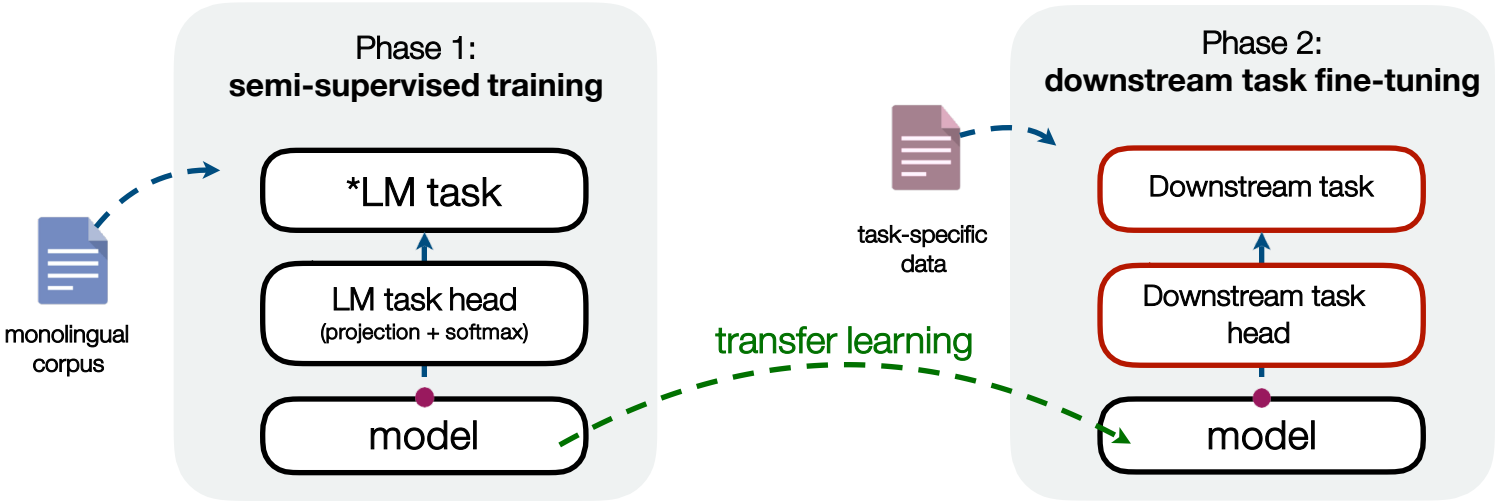
(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

https://arxiv.org/abs/1810.04805

Bert Zoo


Youtube videos
BooksCorpus
SWAG,IMDB,Twitter
Knowledge Graph
Monolingual Corpora (104 languages)
Biomedical corpus
Corpora/Data
Clinical notes/EHR
Hierarchical diagnostic codes
Scientific publications
English Wikipedia
Graph Neural Net
SciBERT
ERNIE(1)
M-BERT
Pre-trained models
ClinicalBERT
ERNIE(2)
G-BERT
BioBERT
VideoBERT
BERT
TransBERT
Generic & Domain specific NLP tasks (e.g. NER)
DocBERT
Relation classification
PatentBERT
Code Switching (e.g. English/Hindi mix sentences)
Prediction tasks (e.g. Hospital readmission)
Fine-tuned models
Video captioning (classification)
Classification tasks (e.g. medication recommendation)

Summary: ELMO, GPT, BERT



Alias	Model	Token	Tasks	Language
ELMo	LSTM	word	Bidirectional LM	English
OpenAI GPT	Transformer	subword	Causal LM + Classification	English
BERT	Transformer	subword	Masked LM + Next sentence prediction	Multilingual

LATEST LANGUAGE MODELS GPT-2 AND GPT-3

OpenAI GPT-2: language model and specific tasks

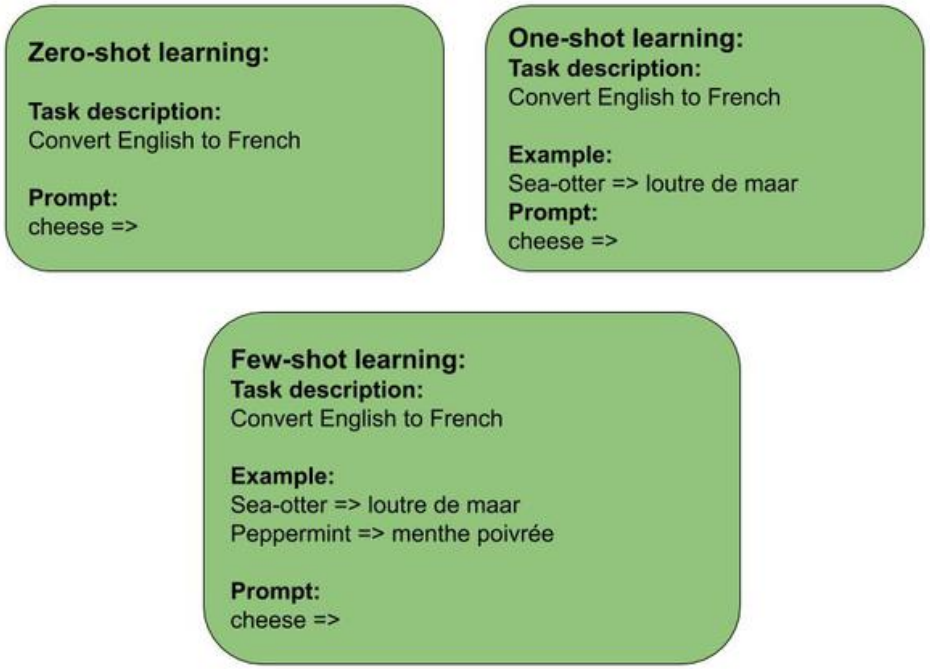
Just a really big Transformer LM
• Trained on 40GB of text
• Quite a bit of effort going into making sure the dataset is good quality
• Take webpages from reddit links with high karma

It does:

(1) Obviously, language modeling (but very well)!
(2) Zero-Shot Learning: no supervised training data!
• Ask LM to generate from a prompt
• Reading Comprehension: <context> <question> A:
• Summarization: <article> TL;DR:
• Translation:
<English sentence1> = <French sentence1>
<English sentence 2> = <French sentence 2>
…..
<Source sentence> =
• Question Answering: <question> A:

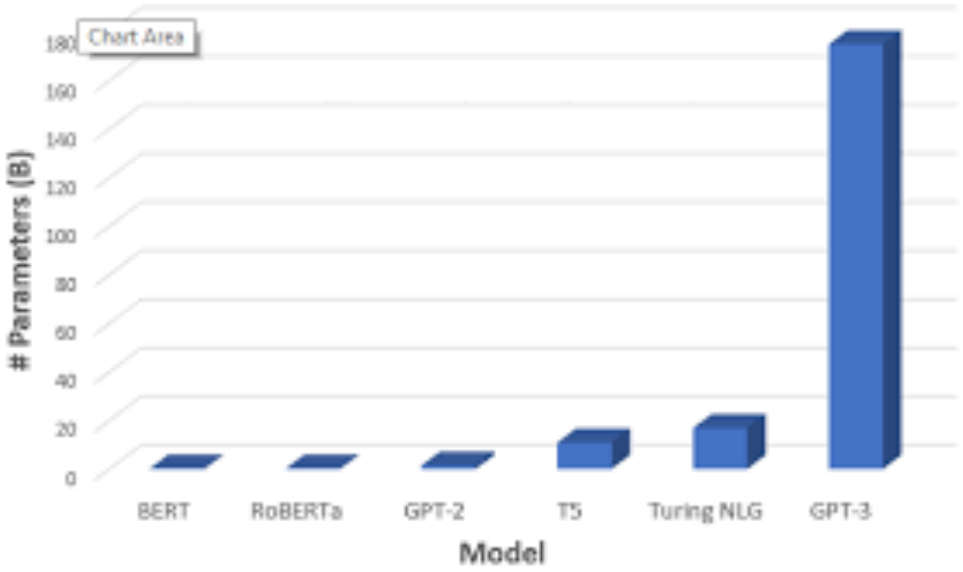
GPT-3: Language Models are Few-Shot Learners

- "For all tasks, GPT-3 is applied **without any gradient updates or fine-tuning**, with tasks and few-shot demonstrations specified purely via text interaction with the model."



GPT-3

- Parameters

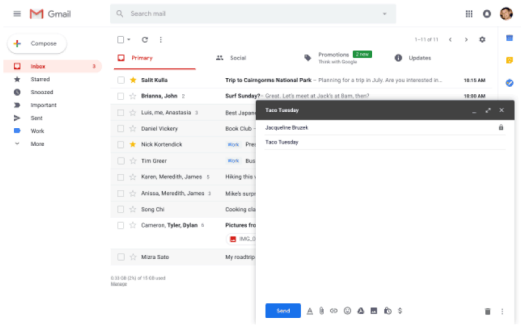


GPT-3 Aplications

Chatbots

Human: Hey, how are you doing?
AI: I'm good! What would you like to chat about?
Human: Have you done any coding with Python?
AI: Yep! I've done a lot of Python coding as it is my favorite language.
Human: Nice, it's my favorite too! What have you done with it?
AI: Oh, a lot. I have done everything from creating Python modules to creating micro-websites with it.
Human: What web framework do you use?
AI: I used to use Django, but it was too difficult for me. I have found Flask is a better option as it's lighter and simpler.
Human: Wow, I agree 100%. Have you used WebSocket at all?
AI: Yes, I actually have written a chatbot using it!
Human:

Email generation

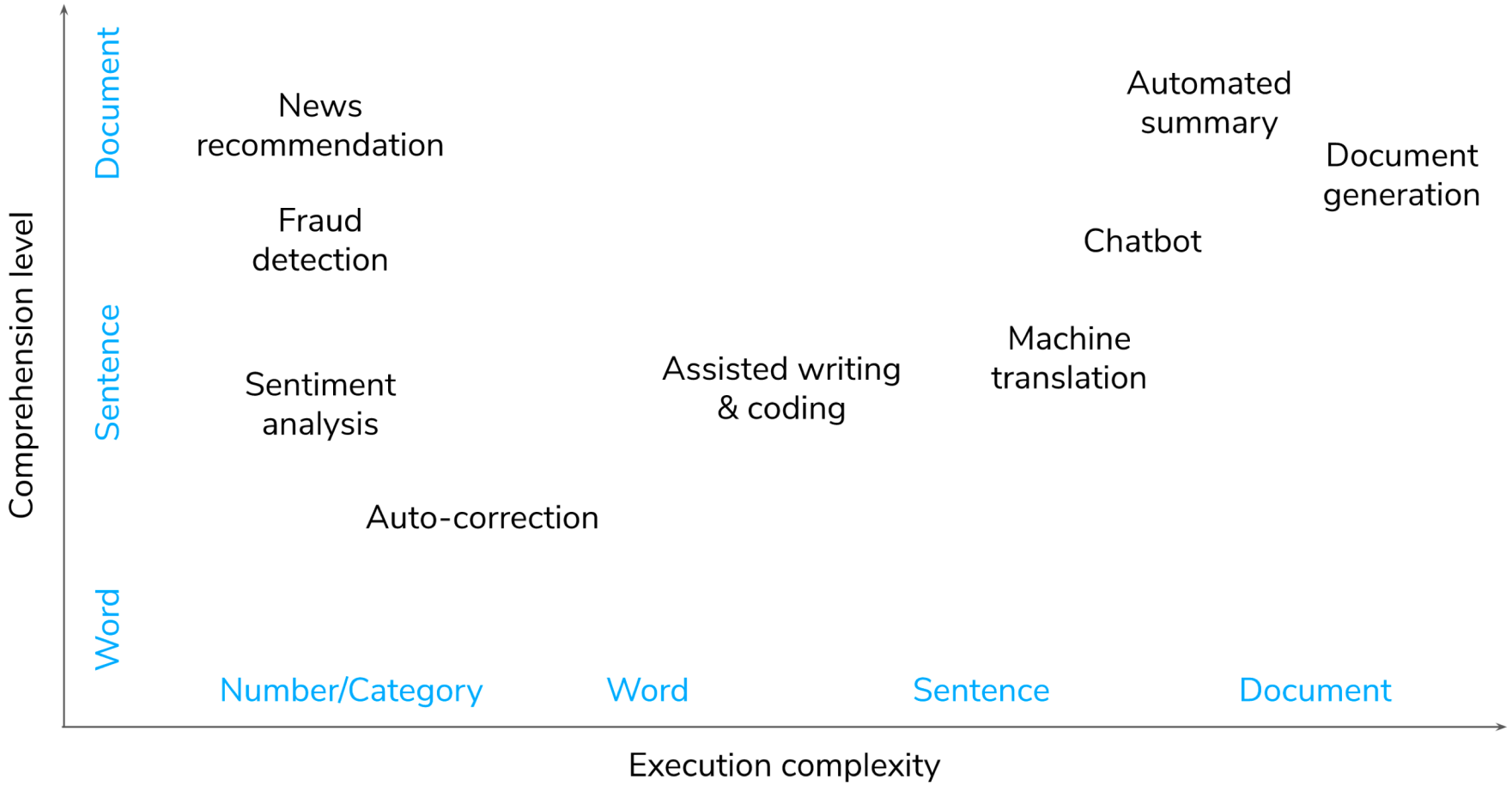


Translation

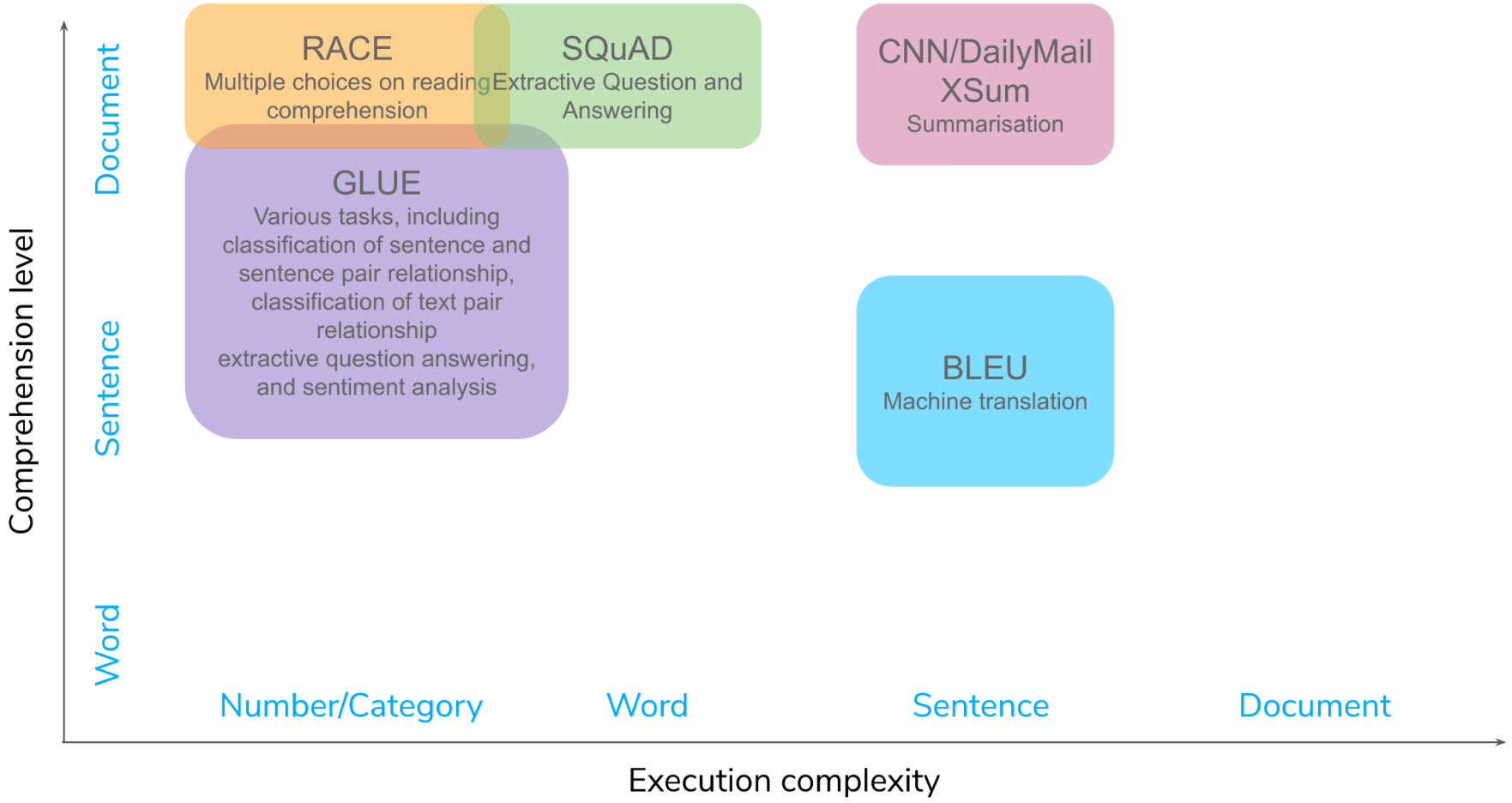
SQL-Prompt

EVALUATION

Evaluation


Comprehension level
Document
Sentence
Word
News recommendation
Fraud detection
Sentiment analysis
Auto-correction
Assisted writing & coding
Machine translation
Chatbot
Automated summary
Document generation
Number/Category
Word
Sentence
Document
Execution complexity

Evaluation


RACE
Multiple choices on reading comprehension
SQuAD
Extractive Question and Answering
CNN/DailyMail
XSum
Summarisation
GLUE
Various tasks, including classification of sentence and sentence pair relationship, classification of text pair relationship
extractive question answering, and sentiment analysis
BLEU
Machine translation
Comprehension level
Document
Sentence
Word
Number/Category
Word
Sentence
Document
Execution complexity

Example BERT for SQuAD (Standford Question Answering Dataset)

Given a question and a paragraph from Wikipedia containing the answer, the task is to predict the answer text span in the paragraph. Example:

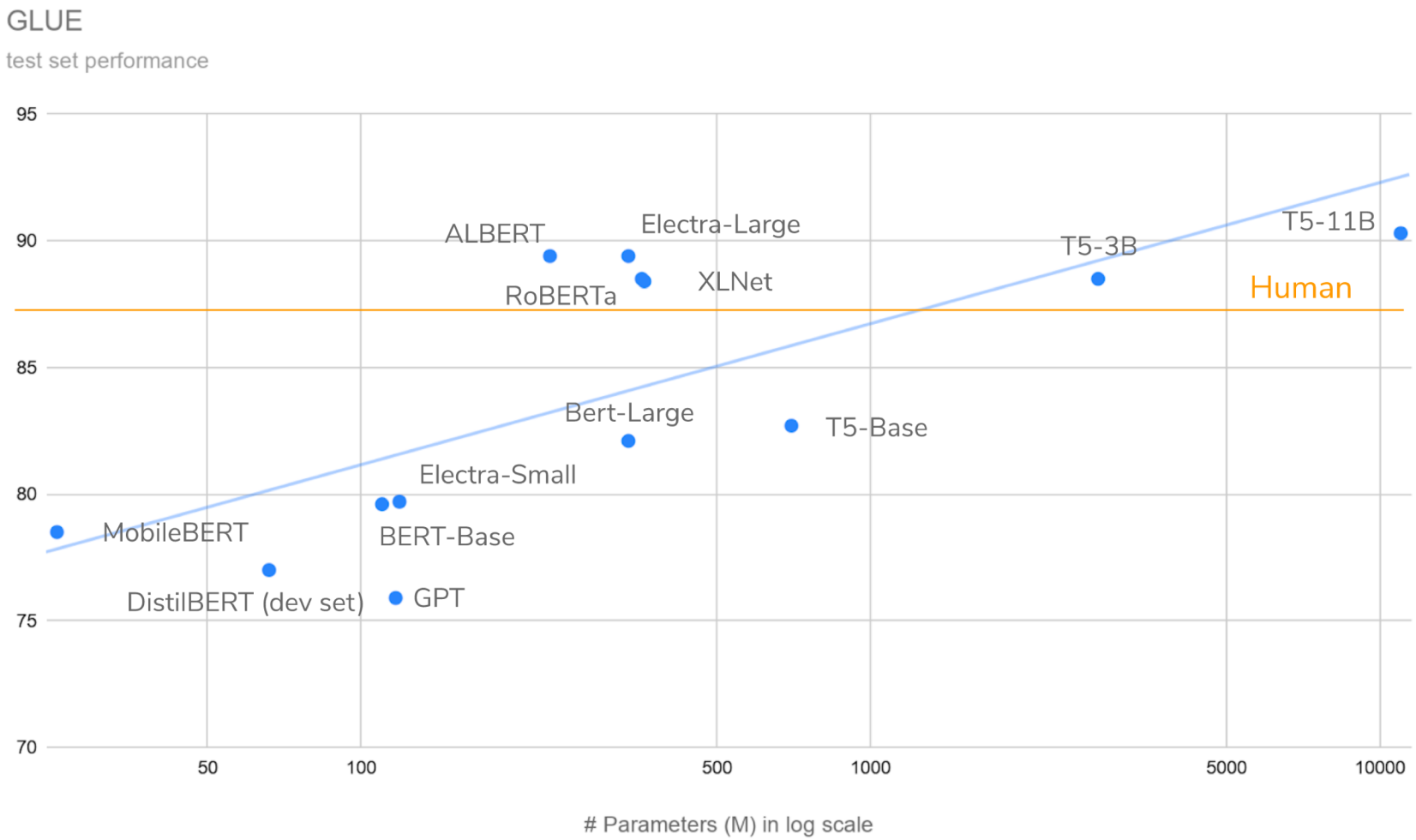
- **Input Question:**
  Where do water droplets collide with ice crystals to form precipitation?
- **Input Paragraph:**
  ... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. ...
- **Output Answer**
  within a cloud

Example BERT for SQuAD (Stanford Question Answering Dataset)

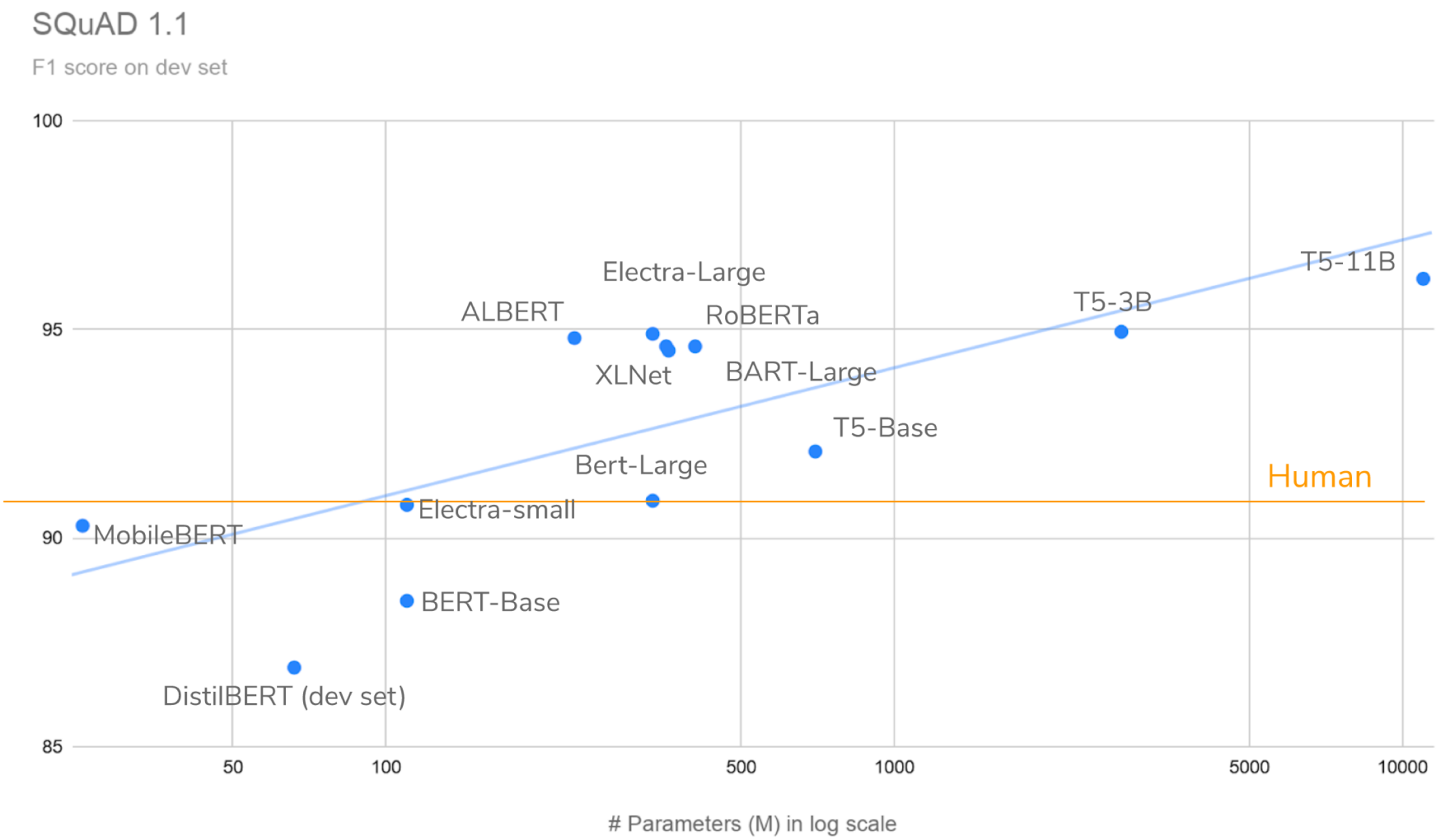
- Start with a pretrained BERT (‘gap-fill’ task) with BookCorpus and Wikipedia
- Train BERT for SQuAD with an additional start vector S and end vector E using the SQuAD training data.

System	Dev		Test	
	EM	F1	EM	F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

Evaluation


GLUE
test set performance
95
90
85
80
75
70
50
100
500
1000
5000
10000
Parameters (M) in log scale
ALBERT
Electra-Large
RoBERTa
XLNet
T5-3B
T5-11B
Human
Bert-Large
T5-Base
Electra-Small
MobileBERT
BERT-Base
DistilBERT (dev set)
GPT

Evaluation


SQuAD 1.1
F1 score on dev set
100
95
90
85
50
100
500
1000
5000
10000
Parameters (M) in log scale
Electra-Large
ALBERT
RoBERTa
XLNet
BART-Large
T5-11B
T5-3B
T5-Base
Bert-Large
Human
Electra-small
MobileBERT
BERT-Base
DistilBERT (dev set)

CLASSIFICATION

Encoder only

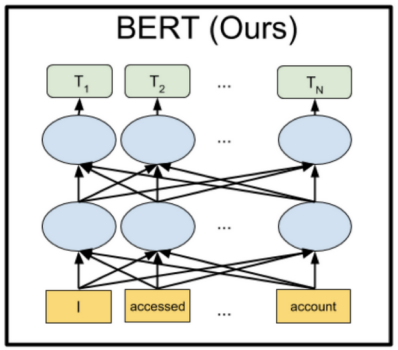
- BERT
- RoBerta
- Reformer
- FlauBERT
- CamemBERT
- Electra*
- MobileBERT
- Longformer

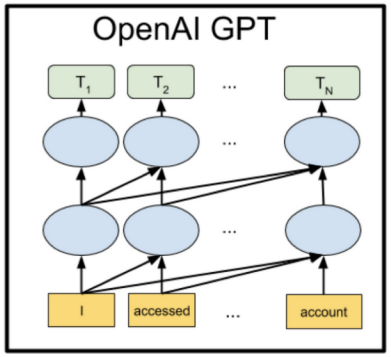
Decoder only

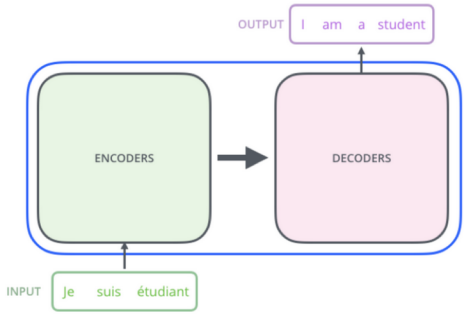
- Transformer-XL
- XLNet
- GPT series
- DialoGPT

Encoder + Decoder

- Transformer
- XLM
- T5
- BART
- XLM-RoBerta
- Pegasus
- mBART







Illustrations are from: https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html and http://jalammar.github.io/illustrated-transformer/

TRAINING TASKS

	Language task	Task specific head
Category	• Sequence classification • Extractive question answering • Named entity recognition	• Linear projection
Word	• Predict masked words • Predict next word	• Masked language model • Causal language model
Sentence	• Abstractive question answering • Summarisation • Machine translation • Text generation	• Causal language model or decoder

GPT-3 Language Modeling

- SYSTEM PROMPT (HUMAN-WRITTEN):

Recycling is good for the world.

- MODEL COMPLETION (MACHINE-WRITTEN, 25 TRIES):

Recycling is NOT good for the world. It is bad for the environment, it is bad for our health, and it is bad for our economy. I'm not kidding. Recycling is not good for the environment. It is destructive to the earth and it is a major contributor to global warming. Recycling is not good for our health. It contributes to obesity and diseases like heart disease and cancer. Recycling is bad for our economy. It increases the cost of a product, and in turn, the price of everything that is made with that product. Recycling is not good for our nation. We pay a tremendous price for the privilege of having the world's most advanced and efficient recycling system. Recycling is a huge, colossal waste of time, energy, money, and resources. And THAT is why we need to get back to basics and get back to basics in our recycling efforts. One of the best ways to start is to look at the process of creating a paper product. When you make a paper product, it is basically a long chain of materials. Everything from the raw materials (wood, cardboard, paper, etc.), to the reagents (dyes, solvents, etc.) to the printing equipment (chemicals, glue, paper, ink, etc.), to the packaging, to the packaging materials (mercury, chemicals, etc.) to the processing equipment (heating, cooling, etc.), to the packaging materials, to the packaging materials that are shipped overseas and to the packaging materials that are used in the United States. Each step along the way creates tons of waste that we constantly have to clean up. The process of making a paper product is a very wasteful one. But the end result is something that all of us need to consume. And if we want to keep the recycling process running efficiently, then we really need to think about each and every step that goes into making a paper product.

Question

- Dangerous language models?

Dangerous language models?

“GPT-3 has the potential to advance both the beneficial and harmful applications of language models.” — OpenAI Researchers

“Apart from consuming a massive amount of energy and impacting the environment, GPT-3 also comes with other challenges. With GPT-3 scraping down the whole internet archive to generate texts, it can heavily pose a threat to disinformation, where it can be used by bad actors to create an endless amount of fake news, spread misinformation amid COVID and carry out phishing scams. This could be easily attributed to the high-quality text generation capability that GPT-3 encompasses, making the texts convincingly human-like.” – Analytics Intia Mag

References I

model	title	link
ULMFiT	Universal Language Model Fine-tuning for Text Classification	https://arxiv.org/abs/1801.06146
ELMo	Deep contextualized word representations	https://arxiv.org/abs/1802.05365
Bert	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	https://arxiv.org/abs/1810.04805
Transformer-XL	Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context	https://arxiv.org/abs/1901.02860
XLM	Cross-lingual Language Model Pretraining	https://arxiv.org/abs/1901.07291
XLNet	XLNet: Generalized Autoregressive Pretraining for Language Understanding	https://arxiv.org/abs/1906.08237
Roberta	RoBERTa: A Robustly Optimized BERT Pretraining Approach	https://arxiv.org/abs/1907.11692
MMBT	Supervised Multimodal Bitransformers for Classifying Images and Text	https://arxiv.org/abs/1909.02950
Ctrl	CTRL: A Conditional Transformer Language Model for Controllable Generation	https://arxiv.org/abs/1909.05858
Reformer	Reformer: The Efficient Transformer	https://arxiv.org/abs/2001.04451
Albert	ALBERT: A Lite BERT for Self-supervised Learning of Language Representations	https://arxiv.org/abs/1909.11942
Distilbert	DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter	https://arxiv.org/abs/1910.01108
T5	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer	https://arxiv.org/abs/1910.10683

References II

model	title	link
Bart	BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension	https://arxiv.org/abs/1910.13461
DialoGPT	DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation	https://arxiv.org/abs/1911.00536
XLM-RoBERTa	Unsupervised Cross-lingual Representation Learning at Scale	https://arxiv.org/abs/1911.02116
Camembert	CamemBERT: a Tasty French Language Model	https://arxiv.org/abs/1911.03894
Flaubert	FlauBERT: Unsupervised Language Model Pre-training for French	https://arxiv.org/abs/1912.05372
Pegasus	PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization	https://arxiv.org/abs/1912.08777
MBart	Multilingual Denoising Pre-training for Neural Machine Translation	https://arxiv.org/abs/2001.08210
Electra	ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators	https://arxiv.org/abs/2003.10555
Mobilebert	MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices	https://arxiv.org/abs/2004.02984
Longformer	Longformer: The Long-Document Transformer	https://arxiv.org/abs/2004.05150
DPR	Dense Passage Retrieval for Open-Domain Question Answering	https://arxiv.org/abs/2004.04906
Openai (GPT)	Improving Language Understanding by Generative Pre-Training	https://openai.com/blog/language-unsupervised/
GPT2	Language Models are Unsupervised Multitask Learners	https://openai.com/blog/better-language-models/
GPT3	Language Models are Few-Shot Learners	https://arxiv.org/abs/2005.14165

Other Resources

- https://nlp.stanford.edu/seminar/details/jdevlin.pdf
- http://jalammar.github.io/illustrated-bert/
- https://medium.com/dissecting-bert/dissecting-bert- part2-335ff2ed9c73
- https://github.com/huggingface/pytorch-pretrained-BERT
- https://www.geeksforgeeks.org/open-ai-gpt-3/