CS182/282A: Designing, Visualizing and Understanding Deep Neural Networks

John Canny

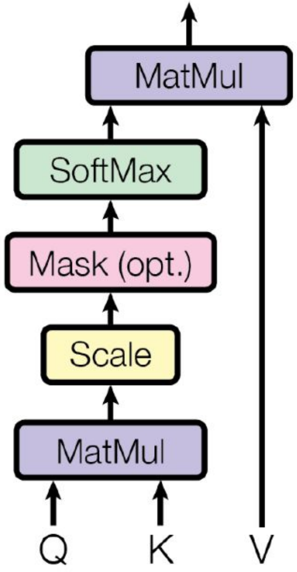
Spring 2020

Lecture 15: NLP Applications with Transformers

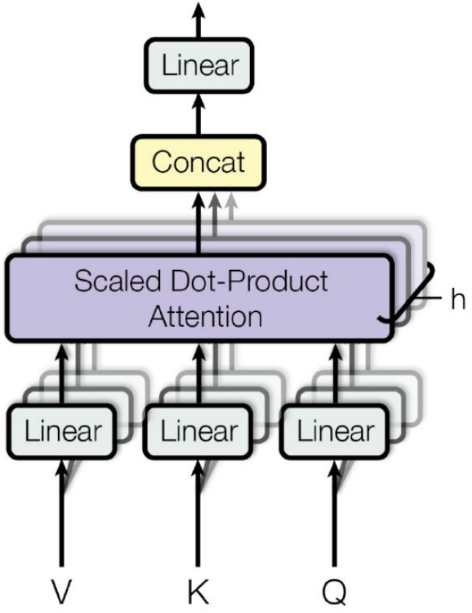
Slides by: Philippe Laban

Last Time - Transformer Attention

QKV Attention
For self and cross attention



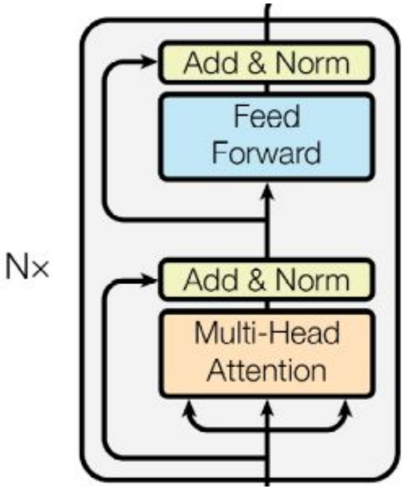
Multi-headed Attention
Several parallel small "heads"
Instead of single bigger one



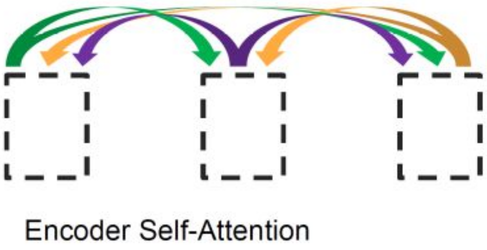
Last Time - Transformer Basic Block

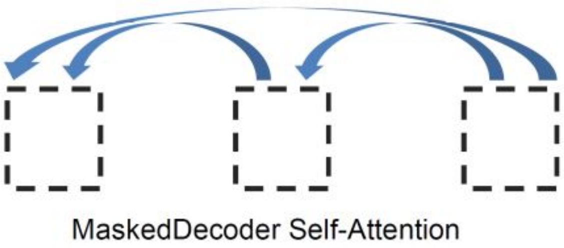
Basic Pattern of a Transformer Block:

[1] Self-attention ("assemble evidence")
([1b] Decoder only: Cross-attention)
[2] Fully-connected ("transform")
[3] Residual layers & normalize
[4] Stack N layers



Last Time - Transformer Attention Masking





Bidirectional attention
Each word attends to words on its "left" and "right" (look at purple arrows)

Used on Transformer Encoders

Cannot be used for generation

Masked attention
A.k.a: *Causal, Left-only, or Decoder* attention

Each word only attends to words appearing previously in the sequence (left)

Used on Transformer Decoders

Used for generation (compatible with auto-regression)

Announcements

We are fully online now!
Just type questions in the zoom chat (click at bottom of screen to get chat window)

Zoom link is: https://berkeley.zoom.us/j/7180580000

Office hours and discussion plans will be posted via piazza and bcourses.

Midterm 2 will be a take-home, currently planned to happen at the same date and time.

Generation vs. Understanding

NLP tasks are often organized into Natural Language **Understanding (NLU)** and Natural Language **Generation (NLG)**

NLG: given text and other inputs, produce (generate) a sequence of text one word at a time, most often in a left-to-right auto-regressive way.

NLU: Given text and other input(s), build a representation of the inputs that can be used for a task.

Generation vs. Understanding

NLP tasks are often organized into Natural Language **Understanding (NLU)** and Natural Language **Generation (NLG)**

NLG: given text and other inputs, produce (generate) a sequence of text one word at a time, most often in a left-to-right auto-regressive way.

NLU: Given text and other input(s), build a representation of the inputs that can be used for a task.

It's a matter of framing: **Summarization**

Generation vs. Understanding

NLP tasks are often organized into Natural Language **Understanding (NLU)** and Natural Language **Generation (NLG)**

NLG: given text and other inputs, produce (generate) a sequence of text one word at a time, most often in a left-to-right auto-regressive way.

NLU: Given text and other input(s), build a representation of the inputs that can be used for a task.

It's a matter of framing: **Summarization**

Abstractive Summarization
Encode the document to be summarized, then generate the summary one word at at time

NLG

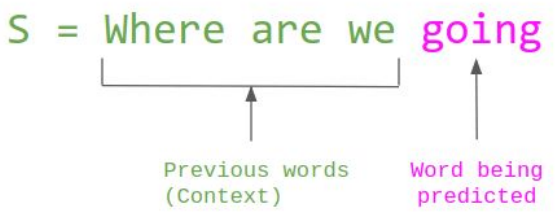
Extractive Summarization
Encode the document to be summarized, select the 3 (or k) most relevant/important sentences.

NLU

Generation vs. Understanding

NLG

Archetype model: **GPT2**
Pretraining: Language Modeling



S = Where are we going

Previous words (Context)

Word being predicted

P(S) = P(Where) x P(are | Where) x P(we | Where are) x P(going | Where are we)

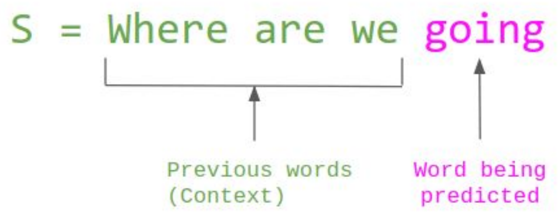
Common Finetuning: Conditional Language modeling (teacher forcing)

(In HW3)

Generation vs. Understanding

NLG

Archetype model: **GPT2**
Pretraining: Language Modeling



S = Where are we going

Previous words (Context)

Word being predicted

P(S) = P(Where) x P(are | Where) x P(we | Where are) x P(going | Where are we)

Common Finetuning: Conditional Language modeling (teacher forcing)

(In HW3)

NLU

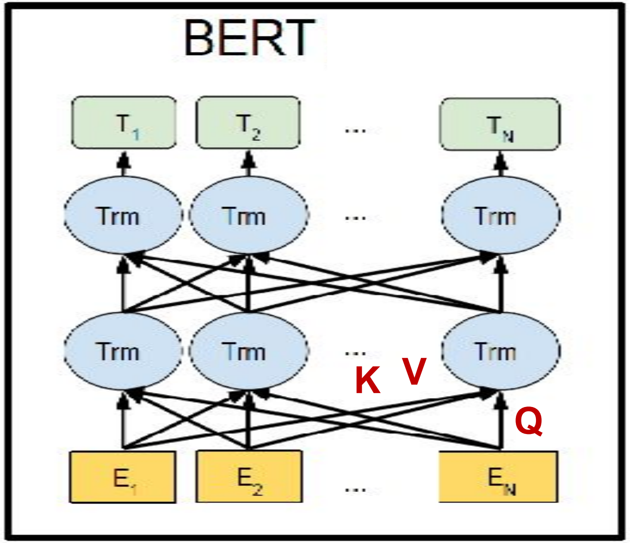
Archetype model: **BERT**
Pretraining: Masked Language Modeling + Is Next

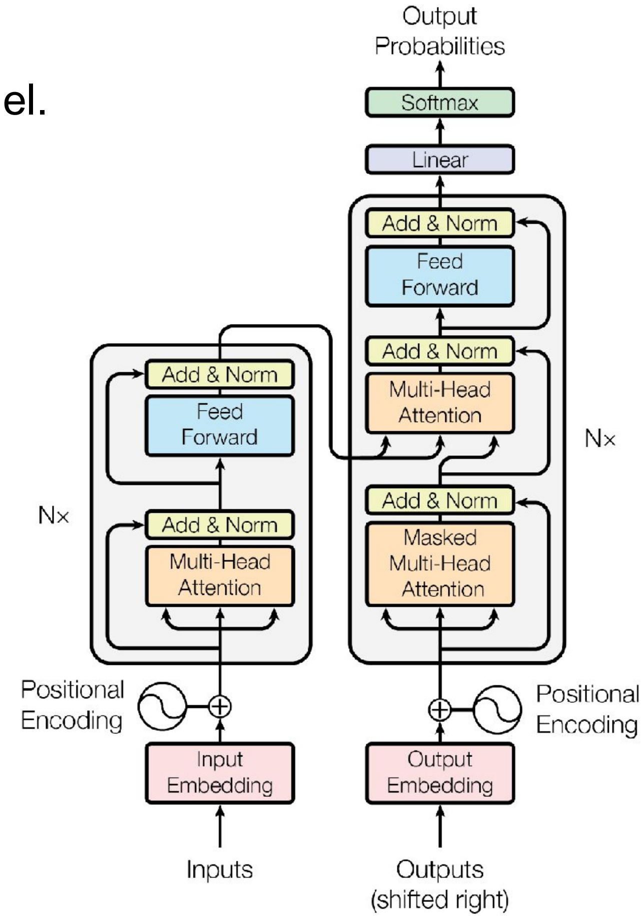
Finetuning: Task dependent (4+ types)

(NLP Project)

BERT

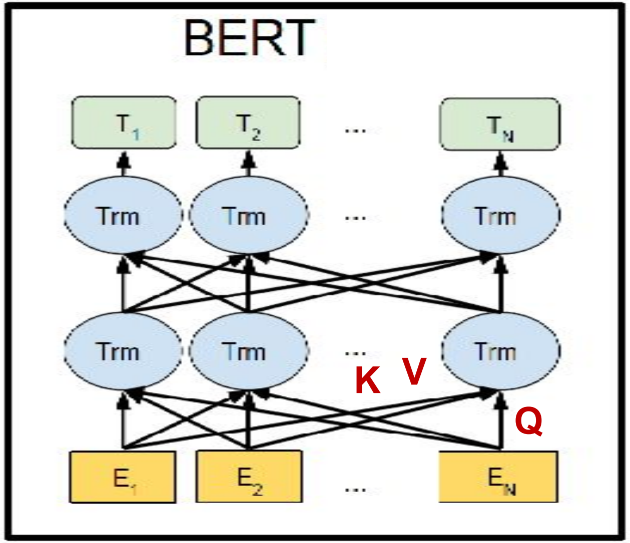
BERT is a bidirectional (Transformer encoder) model.

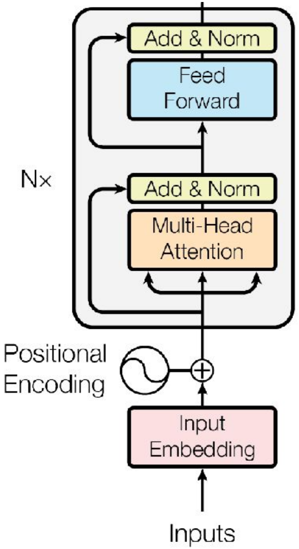




BERT

BERT is a bidirectional (Transformer encoder) model.





BERT

BERT is trained with two types of loss:

A: Masked Language Modeling (MLM)

- Select 15% random words in the input.
- Replace each by a <MASK> token
- On the output re-predict the original words
- There is a bit more detail, see detail in the paper

BERT

BERT is trained with two types of loss:

A: Masked Language Modeling (MLM)

- Select 15% random words in the input.
- Replace each by a <MASK> token
- On the output re-predict the original words

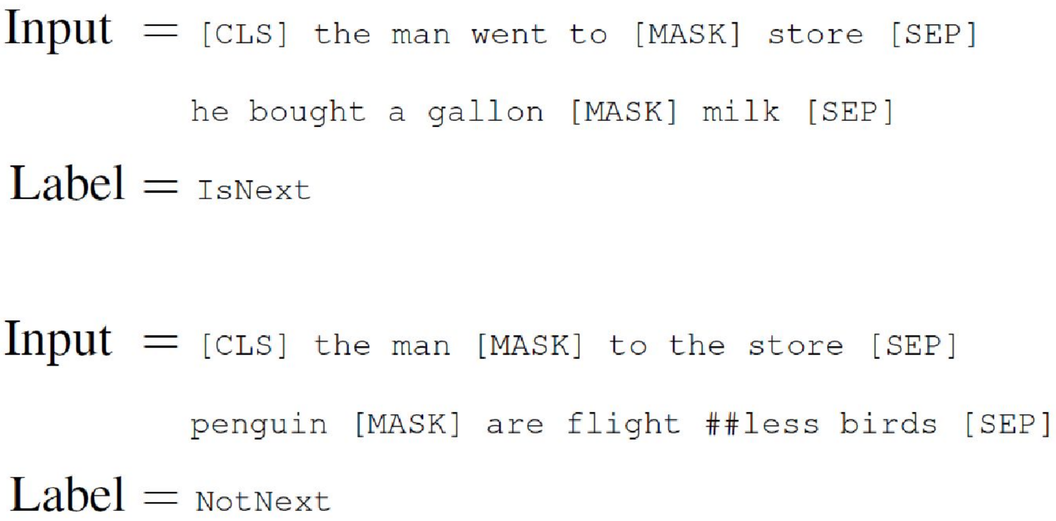
B: Next sentence prediction.

- Given a corpus of consecutive sentence pairs (text documents)
- Create a dataset with 50% real pairs of consecutive sentences, and 50% “fake” pairs (where the second sentence is a random one).
- Build inputs of the form: *<CLS> Sentence 1 <SEP> Sentence 2 <SEP>*
- Use CLS output to classify pair as real or fake consecutive pair

This is a *contrastive loss* (common in unsupervised learning). Contrasting true positives with “near miss” negatives.

BERT

BERT is trained with two types of loss:



Input = [CLS] the man went to [MASK] store [SEP]

he bought a gallon [MASK] milk [SEP]

Label = IsNext

Input = [CLS] the man [MASK] to the store [SEP]

penguin [MASK] are flight ##less birds [SEP]

Label = NotNext

During pretraining, the model must learn to recover what has been MASKed and predict sentence consecutiveness.

Masked vs. Regular Language modeling

MLM and LM seem very similar. Does bi-directionality make a difference in practice?

Yes, bi-directionality is important. When pre-training models of similar size, on similar data:

- The generative model might achieve 50% accuracy at LM
- The bi-directional model might achieve ~65% accuracy at MLM

better pretraining >> better embeddings >> better model when finetuning

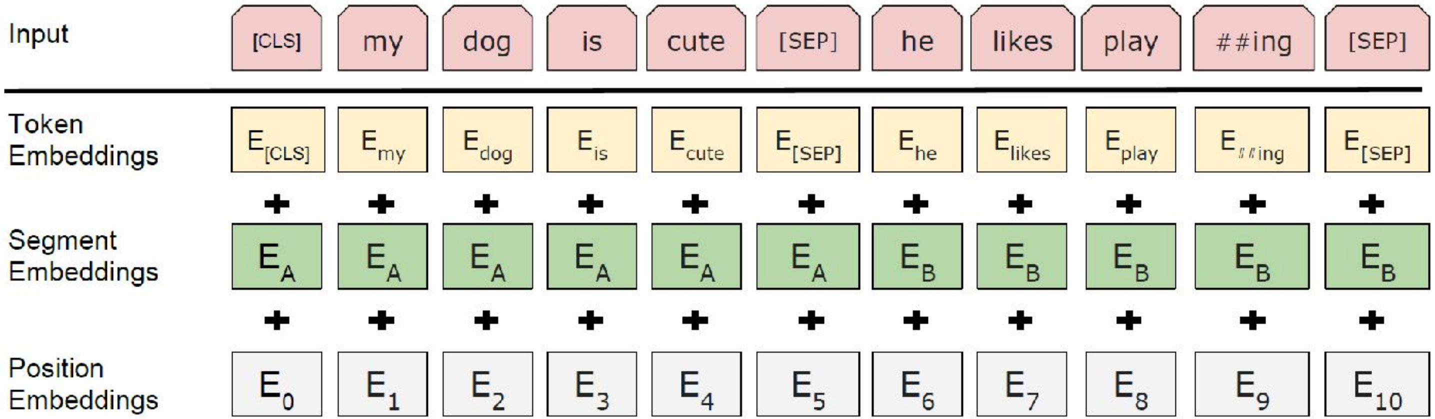
What is a... contextual word embedding

Transformer-based models start with a static word embedding (input of layer 1), which is "*contextualized*" by going through stacked layers of self-attention and transformation.

I run my house's kitchen, so if we run out of something, I do a quick run to the market.

In a Transformer, the model will learn to contextualize and differentiate each instance of "run". In comparison, word2vec has a single vector for "run".

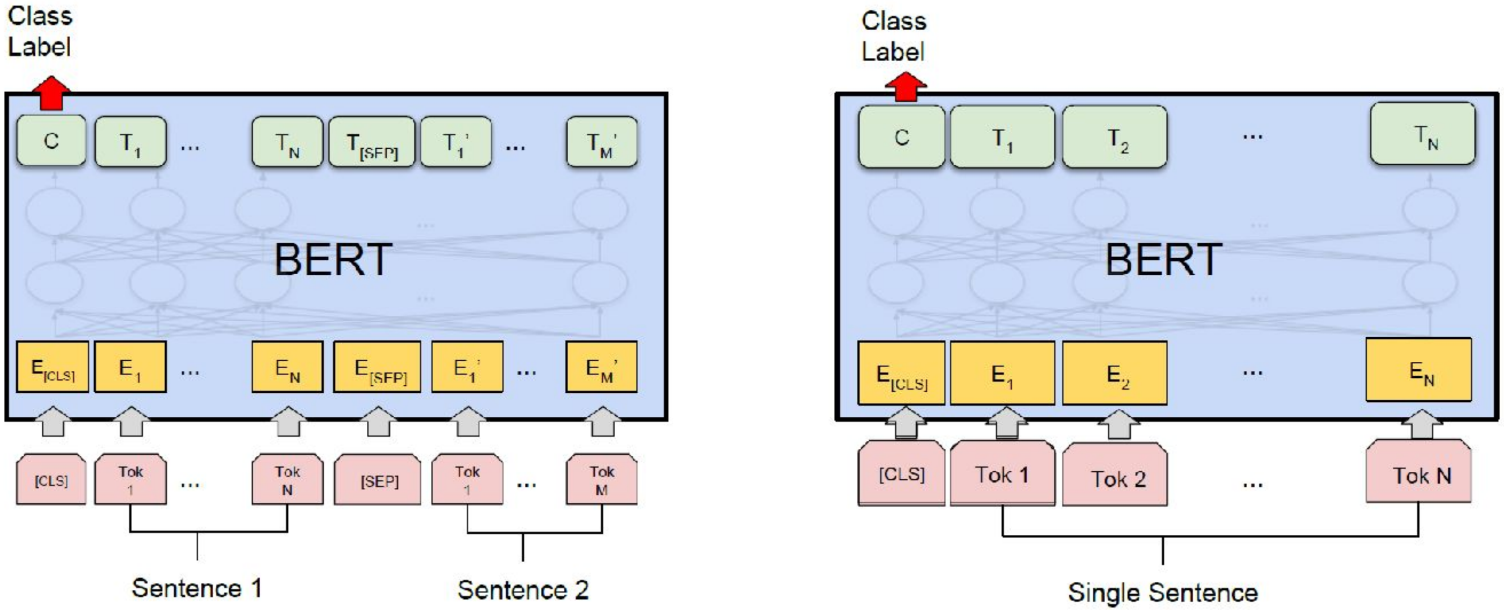
BERT Task specialization



Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{likes}	E_{play}	$E_{\#\#ing}$	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+	+
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
	+	+	+	+	+	+	+	+	+	+	+
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}

Figure 2: BERT input representation. The input embeddings is the sum of the token embeddings, the segmentation embeddings and the position embeddings.

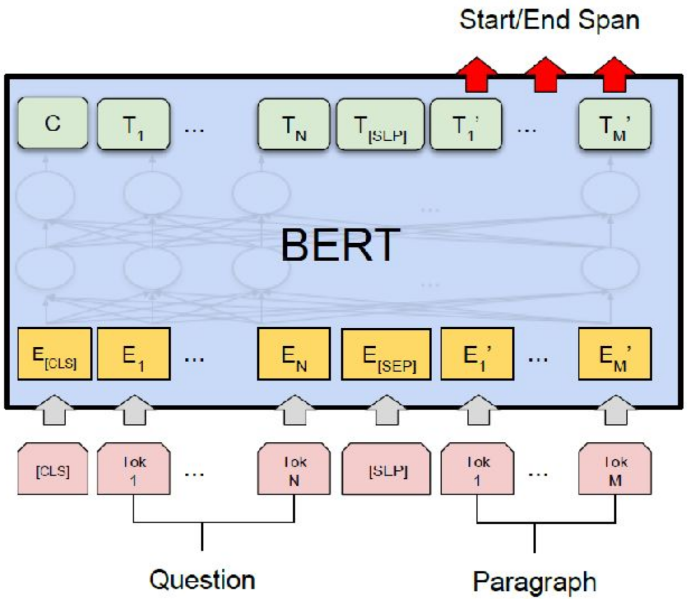
BERT Task specialization



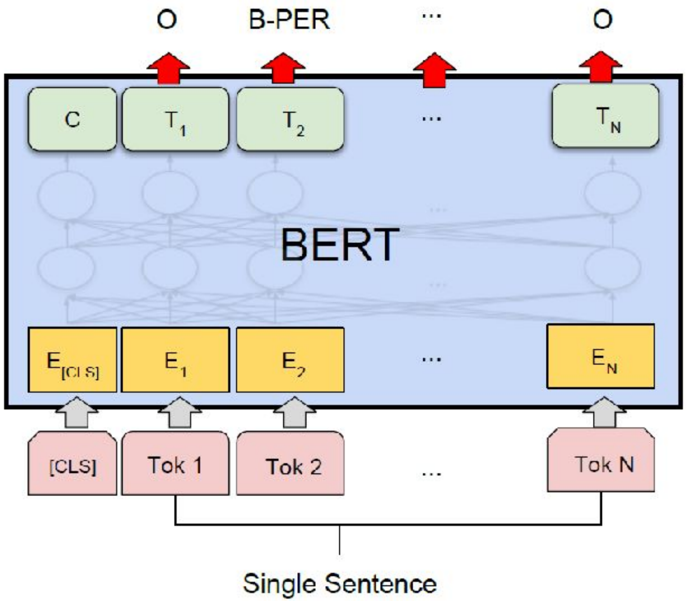
(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG

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

BERT Task specialization



(c) Question Answering Tasks:
SQuAD v1.1



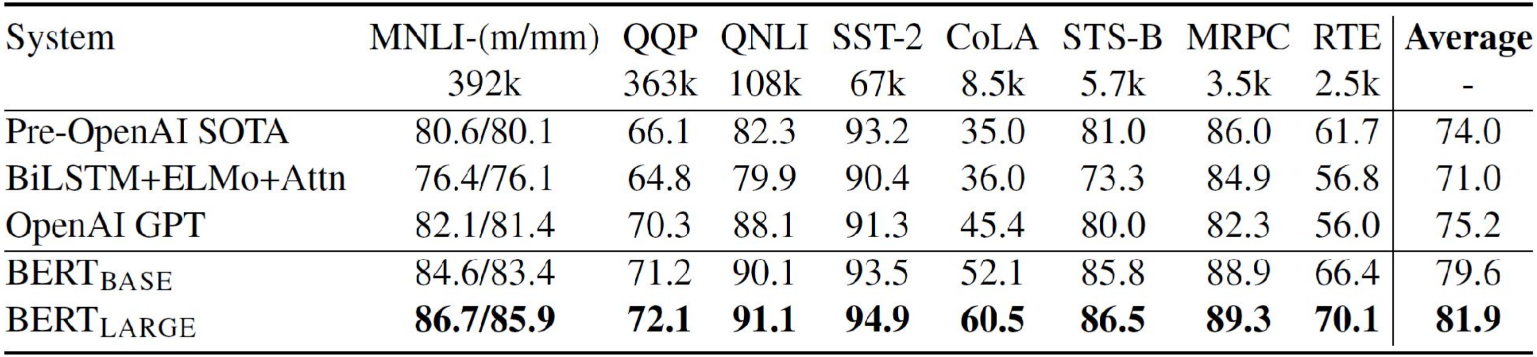
(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER

BERT Performance

BERT Base: L=12, H=768, A=12, total param =110M
BERT Large: L=24, H=1024, A=16, total param=340M

L=number of layers, H=model dimension, A=number of multi-attention heads

General Language Understanding Evaluation (GLUE) tasks



System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

Natural Language Inference

NLI is also referred to as "entailment".

Given a premise (P) decide whether a hypothesis (H) is **entailed, contradicts** or is **neutral**. This task is a three-way sentence pair classification..

If we have a premise: P = "Kids are playing soccer in the garden."
And several hypotheses:
H1 = "Someone is in the garden" (**entail)**
H2 = "It is sunny outside" (**neutral**)
H3 = "No one ever goes to the garden" (**contradict**)

Natural Language Inference

If we have a premise: P = "Kids are playing soccer in the garden."

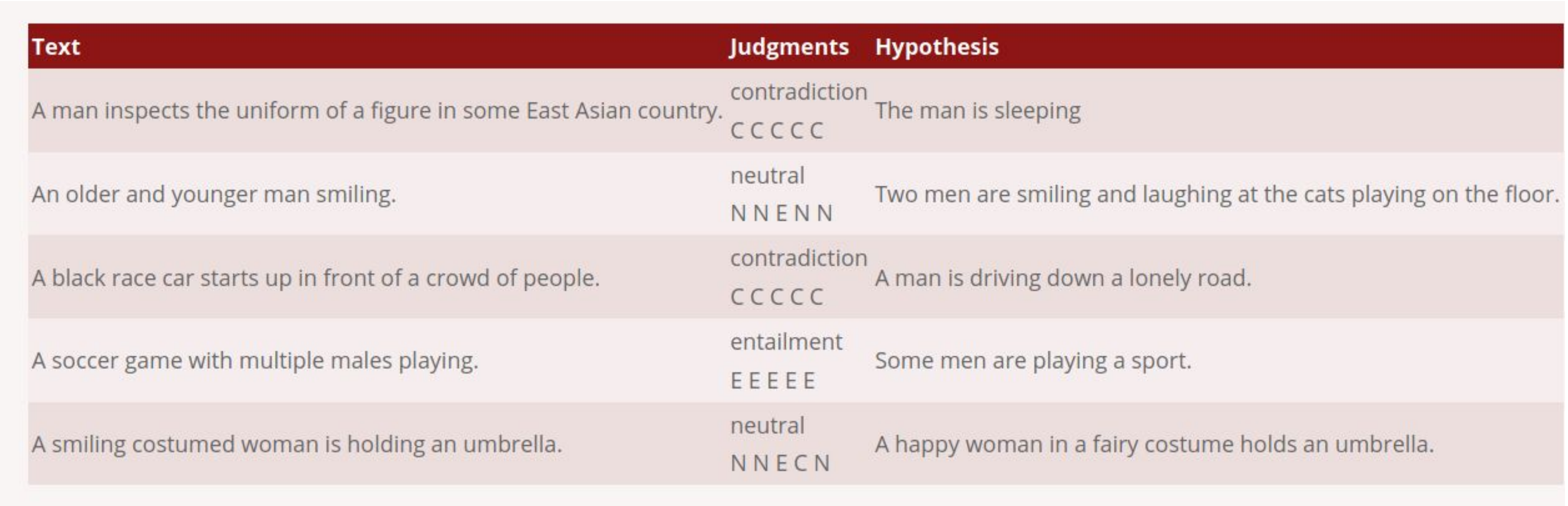
What about:
H4 = "The kids enjoy playing soccer." (??)
H5 = "No adults are playing soccer in the garden." (??)

NLI requires interpretation.
When collecting data, it is important:

- Collect several opinions to measure *inter-annotator agreement*
- Deal with/Remove cases that are ambiguous

Natural Language Inference - Datasets

Stanford Natural Language Inference (570k labeled sentence pairs)
Based on image captions from Flickr



Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

Natural Language Inference - Datasets

Stanford Natural Language Inference (SNLI) (570k labeled sentence pairs)
Based on image captions from Flickr

Limitation: Entailment in Image captions is very specific.

Next iteration of dataset: MultiGenre NLI (MNLI)
The dataset is made from 10 different genres of text: Fiction, Letters, Telephone speech, etc.

Challenging dataset: part of GLUE.

Language Acceptability

Corpus of Language Acceptability (CoLA) measure a model to recognize syntax correctness (acceptability).

In which way is Sandy very anxious to see if the students will be able to solve the homework problem? **Label: 0**

The book was written by John. **Label: 1**

Books were sent to each other by the students. **Label: 0**

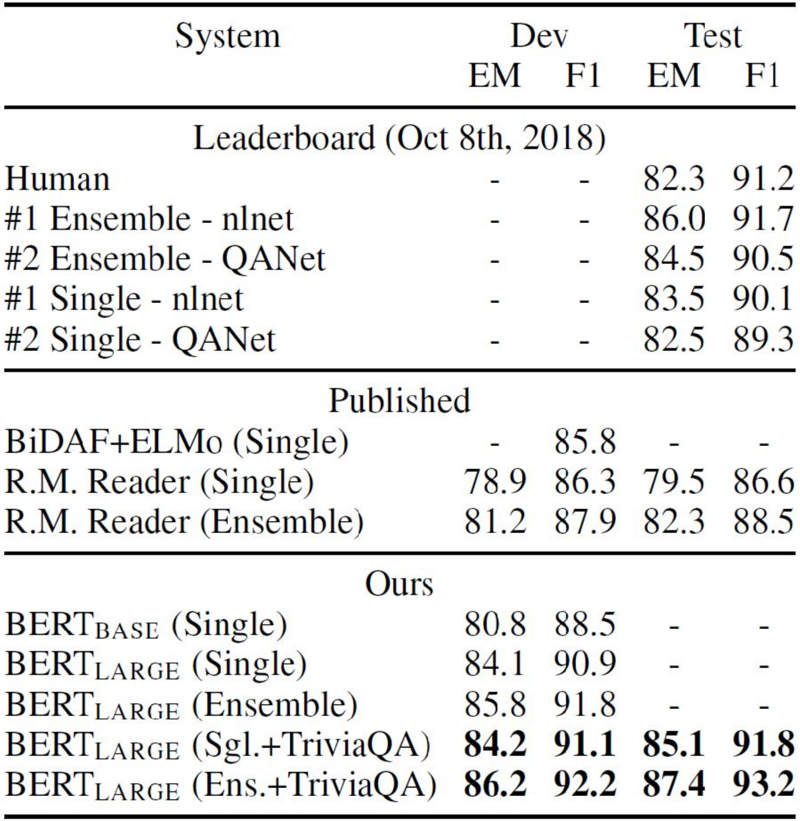
She voted for herself. **Label: 1**

I saw that gas can explode. **Label: 0**

NLI is a semantic task. CoLA is a syntactic task, focusing on grammar.

BERT Performance

SQuAD

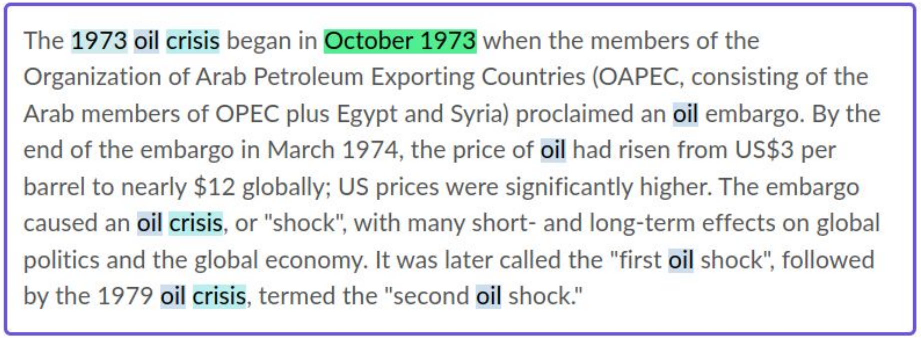


System	Dev		Test	
	EM	F1	EM	F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
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**

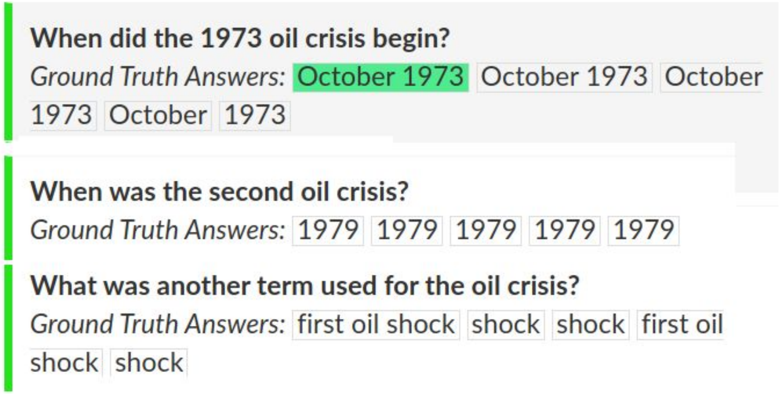
Question Answering

SQuAD 1.0: Stanford Question Answering Dataset

Given a paragraph P, a question Q, find the answer as a subset of P (start and end character)



The 1973 oil crisis began in October 1973 when the members of the Organization of Arab Petroleum Exporting Countries (OAPEC, consisting of the Arab members of OPEC plus Egypt and Syria) proclaimed an oil embargo. By the end of the embargo in March 1974, the price of oil had risen from US$3 per barrel to nearly $12 globally; US prices were significantly higher. The embargo caused an oil crisis, or "shock", with many short- and long-term effects on global politics and the global economy. It was later called the "first oil shock", followed by the 1979 oil crisis, termed the "second oil shock."



When did the 1973 oil crisis begin?
Ground Truth Answers: October 1973 | October 1973 | October 1973 | October 1973

When was the second oil crisis?
Ground Truth Answers: 1979 | 1979 | 1979 | 1979 | 1979

What was another term used for the oil crisis?
Ground Truth Answers: first oil shock | shock | shock | first oil shock | shock

Question Answering

SQuAD 1.0/1.1: **S**tanford **Qu**estion **A**nswering **D**ataset

Given a paragraph P, a question Q, find the answer as a subset of P (start and end character)

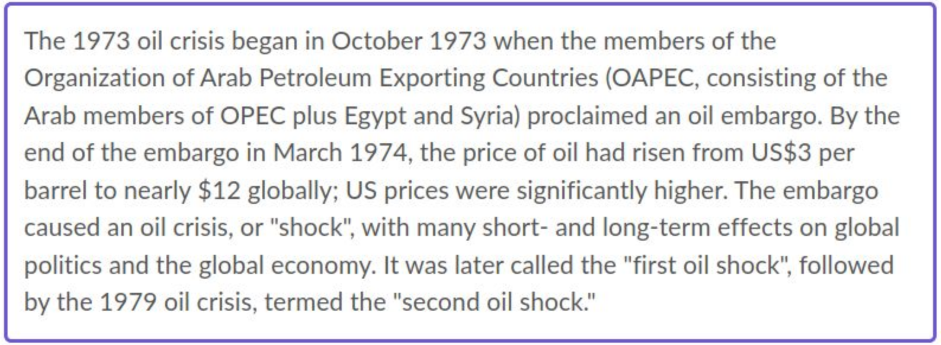
Limitation: Answer is always in the paragraph. This is a big bias. If you ask a "When ..." question and there is a single date in the text, it must be the answer. Not usable in practice, too many false positives.

Question Answering

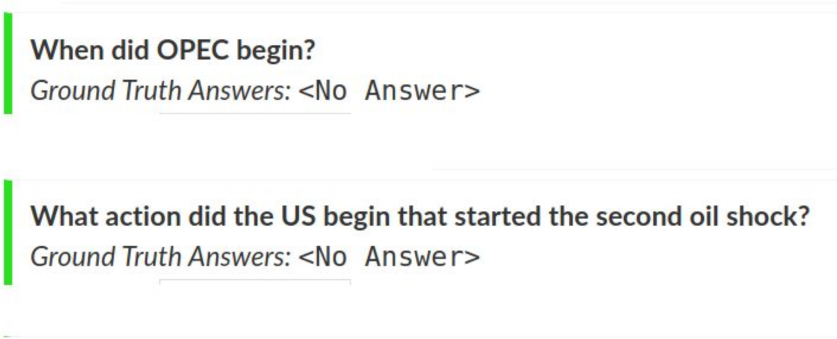
SQuAD 2.0: Stanford Question Answering Dataset

Given a paragraph P, a question Q:

- Decide whether the answer is in the paragraph,
- If it is, extract the subset of the paragraph answering the question



The 1973 oil crisis began in October 1973 when the members of the Organization of Arab Petroleum Exporting Countries (OAPEC, consisting of the Arab members of OPEC plus Egypt and Syria) proclaimed an oil embargo. By the end of the embargo in March 1974, the price of oil had risen from US$3 per barrel to nearly $12 globally; US prices were significantly higher. The embargo caused an oil crisis, or "shock", with many short- and long-term effects on global politics and the global economy. It was later called the "first oil shock", followed by the 1979 oil crisis, termed the "second oil shock."



When did OPEC begin?

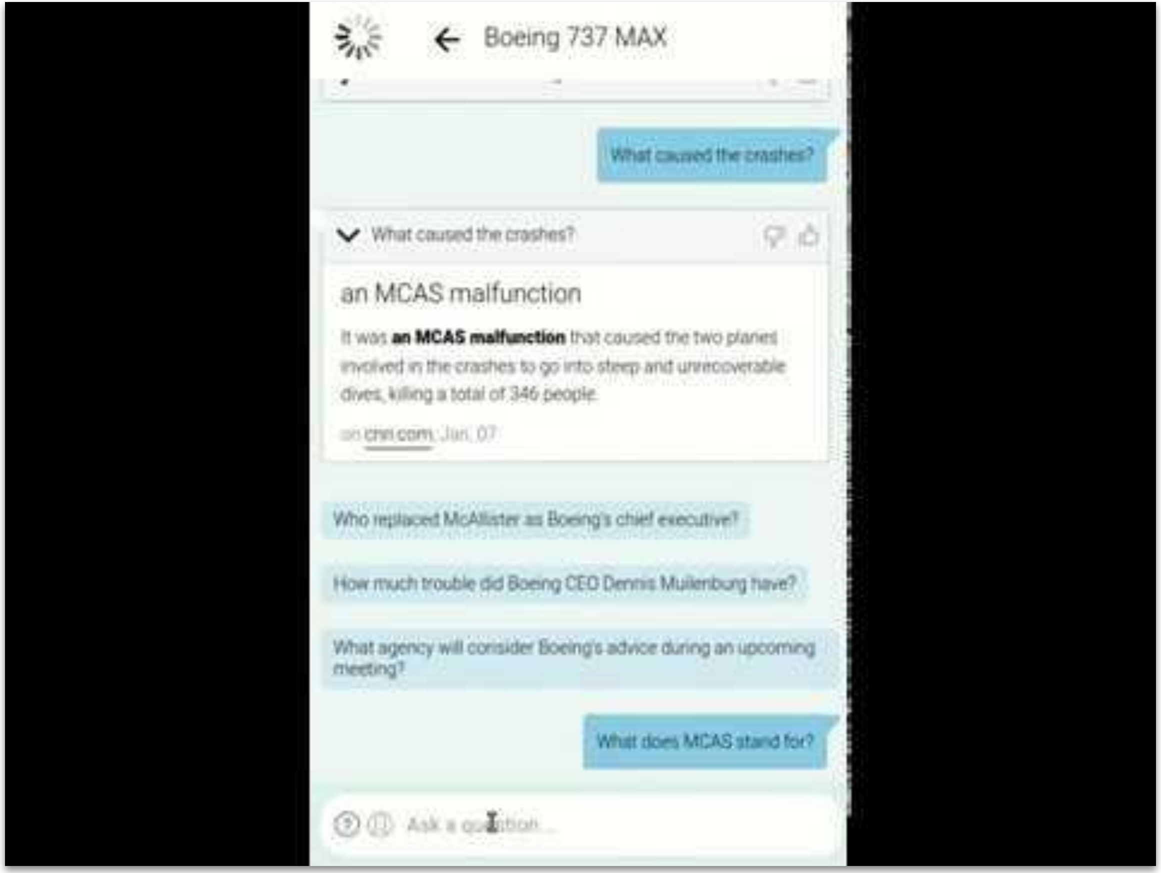
Ground Truth Answers: <No Answer>

What action did the US begin that started the second oil shock?

Ground Truth Answers: <No Answer>

More difficult task, more usable in practice.

Q&A Demo



What caused the crashes?

What caused the crashes?

an MCAS malfunction

It was **an MCAS malfunction** that caused the two planes involved in the crashes to go into steep and unrecoverable dives, killing a total of 346 people.

on cnn.com, Jan. 07

Who replaced McAllister as Boeing's chief executive?

How much trouble did Boeing CEO Dennis Muilenburg have?

What agency will consider Boeing's advice during an upcoming meeting?

What does MCAS stand for?

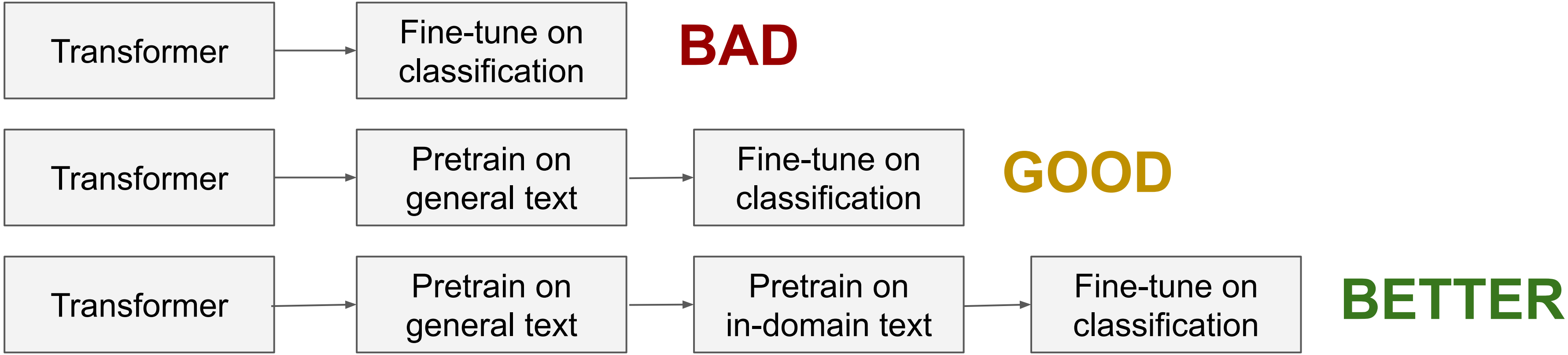
Demo of a research project using a Q&A system on news data in Prof. Canny's lab

Training a classifier with Transformers

Practical tip:

When training a classifier, pretrain the model on data as close to in-domain data.

For in detail analysis: ULMFit paper 2018



Transformer → Fine-tune on classification **BAD**

Transformer → Pretrain on general text → Fine-tune on classification **GOOD**

Transformer → Pretrain on general text → Pretrain on in-domain text → Fine-tune on classification **BETTER**

Back to Generation

Photo by: Daniel McCullough

Evaluating Language Models - Perplexity

The Perplexity metric in NLP is a way to capture the degree of 'uncertainty' a model has in predicting (assigning probabilities to) some text. Lower the entropy (uncertainty), lower the perplexity.

Two language models can be compared by calculating their perplexity on a common test set.

$$\begin{aligned} PP(W) &= P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}} \\ &= \sqrt[N]{\frac{1}{P(w_1 w_2 \ldots w_N}}) \end{aligned}$$

$$PP(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_1 \ldots w_{i-1})}}$$

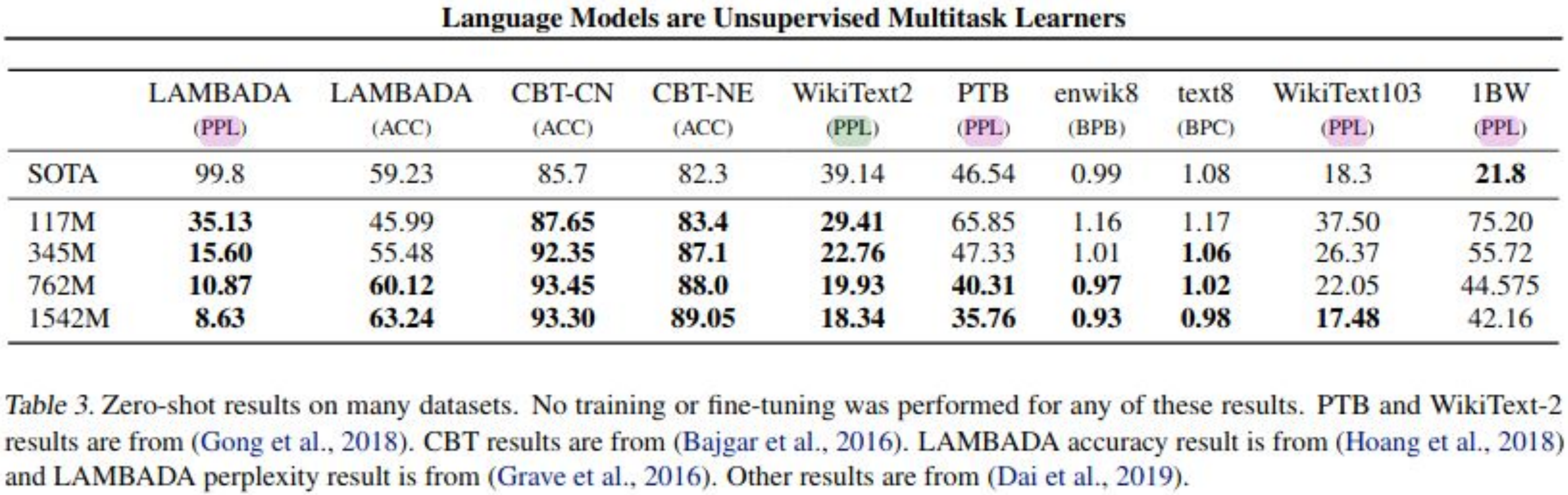
Evaluating Language Models - Perplexity

The Perplexity metric in NLP is a way to capture the degree of 'uncertainty' a model has in predicting (assigning probabilities to) some text. Lower the entropy (uncertainty), lower the perplexity.

I've heard I can compute perplexity as the exponential of an LM's cross-entropy. Is that true?

That's correct, as long as the perplexities being compared use the **same vocabulary**. When model vocabularies differ, normalization must be applied (e.g. perplexity per character). Some references in the additional material.

Perplexity of Transformer Models - GPT2



Language Models are Unsupervised Multitask Learners

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

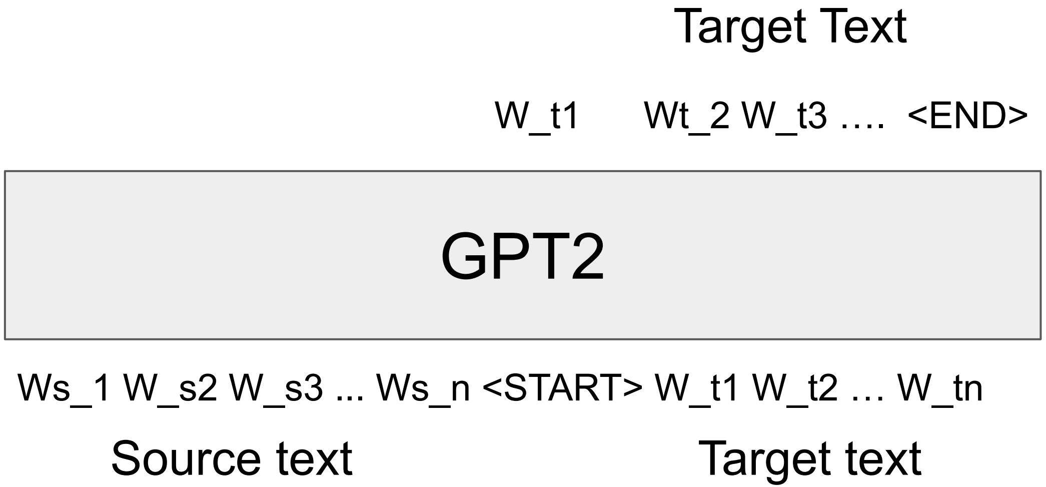
Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

Remarkable fact: GPT2 outperformed state-of-the-art in 7 out of 8 common Language Modeling datasets, **without finetuning.** Figure from GPT2 paper.

How to use GPT2 model

Many generation tasks can be framed as conditional language modeling:

Summarization = P(summary | document)
Dialog Agent = P(system response | dialog history)
Question Generation = P(question | paragraph)
Translation = P(target sentence|source sentence)



Target Text

W_t1 Wt_2 W_t3 …. <END>

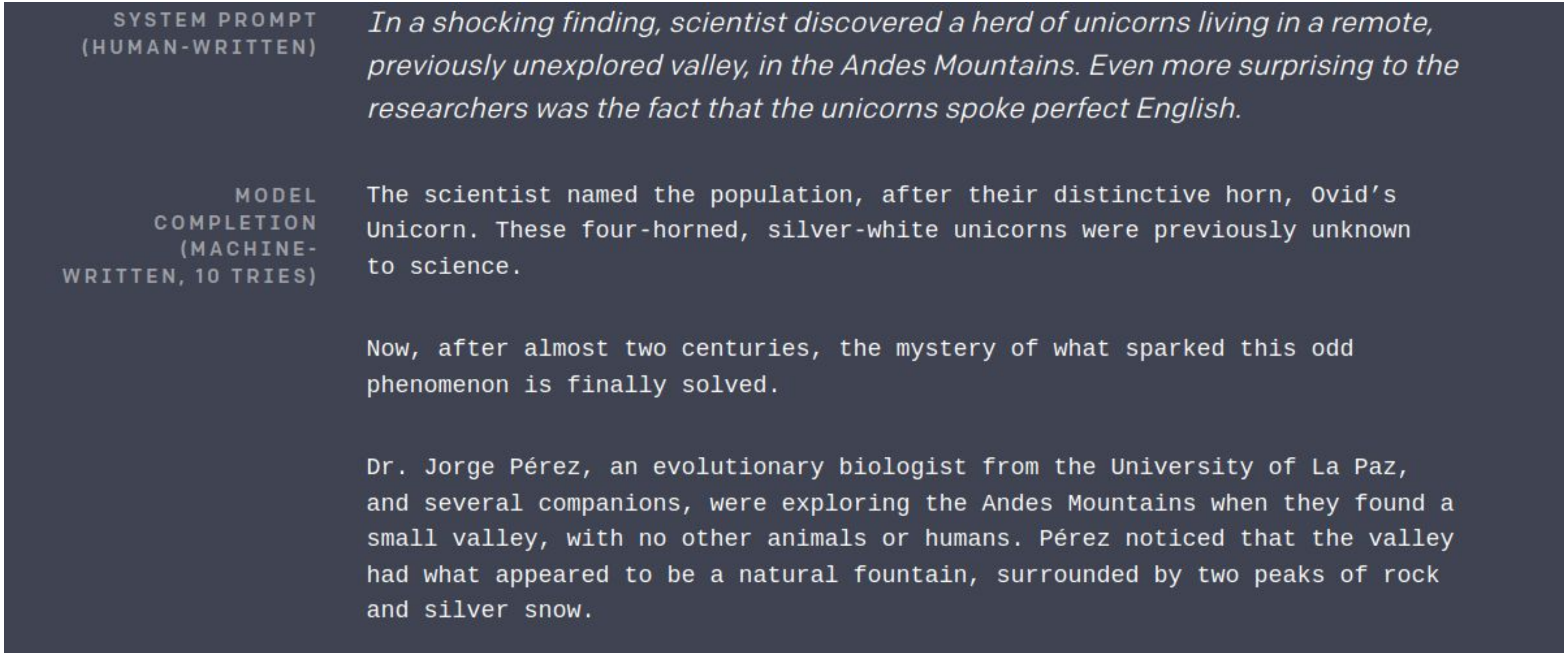
GPT2

Ws_1 W_s2 W_s3 ... Ws_n <START> W_t1 W_t2 … W_tn

Source text

Target text

GPT2 Conditional Generation at work



SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

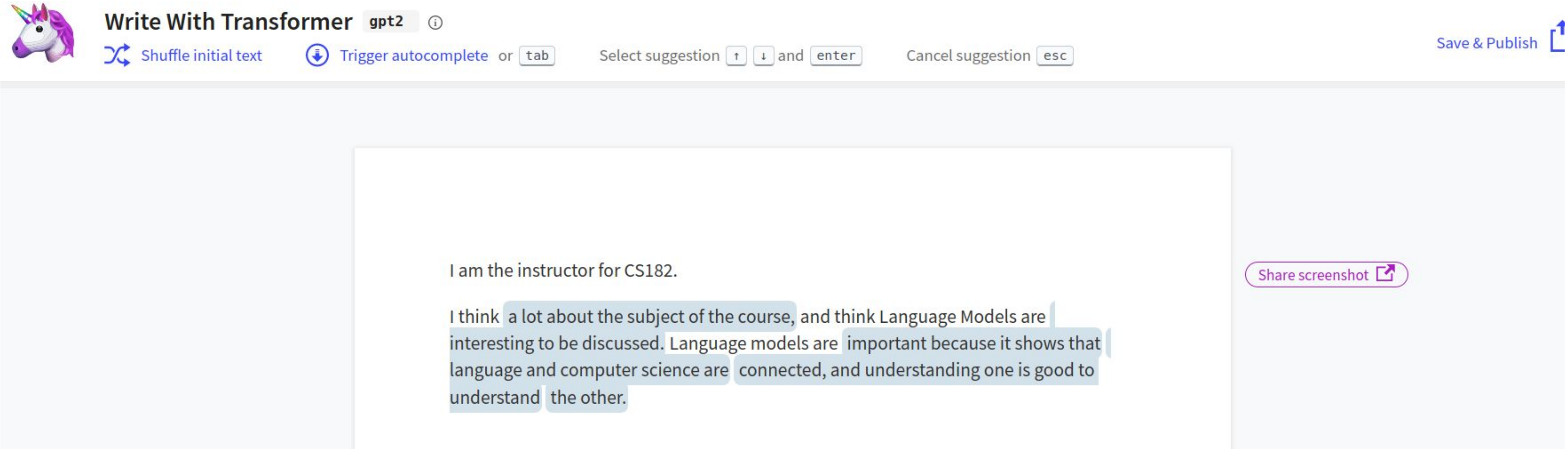
The scientist named the population, after their distinctive horn, Ovid’s Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

From the GPT2 blog post. The model that generated this sample was at first not release, out of fear it could be used to generate fake news, and other malicious content. It has since been released.

Trying conditional Generation



Write With Transformer gpt2 ⓘ
Shuffle initial text
Trigger autocomplete or tab
Select suggestion ↑ ↓ and enter
Cancel suggestion esc
Save & Publish
Share screenshot

I am the instructor for CS182.

I think a lot about the subject of the course, and think Language Models are interesting to be discussed. Language models are important because it shows that language and computer science are connected, and understanding one is good to understand the other.

HuggingFace has built a demo interface to test out GPT2-large at autocompletion. You can test the difference between GPT2-base and GPT2-large.

Can I train an LM on multiple languages?

Would a single model learn to recognize languages, model them jointly, and learn a common "universal" embedding?

The XLM paper experiments with a "Cross-lingual Language model" (Lample et al. 2019) trained on up to **15 languages.**

They introduce a new task called **Translation Language Modeling**.

TLM - Translation Language Modeling

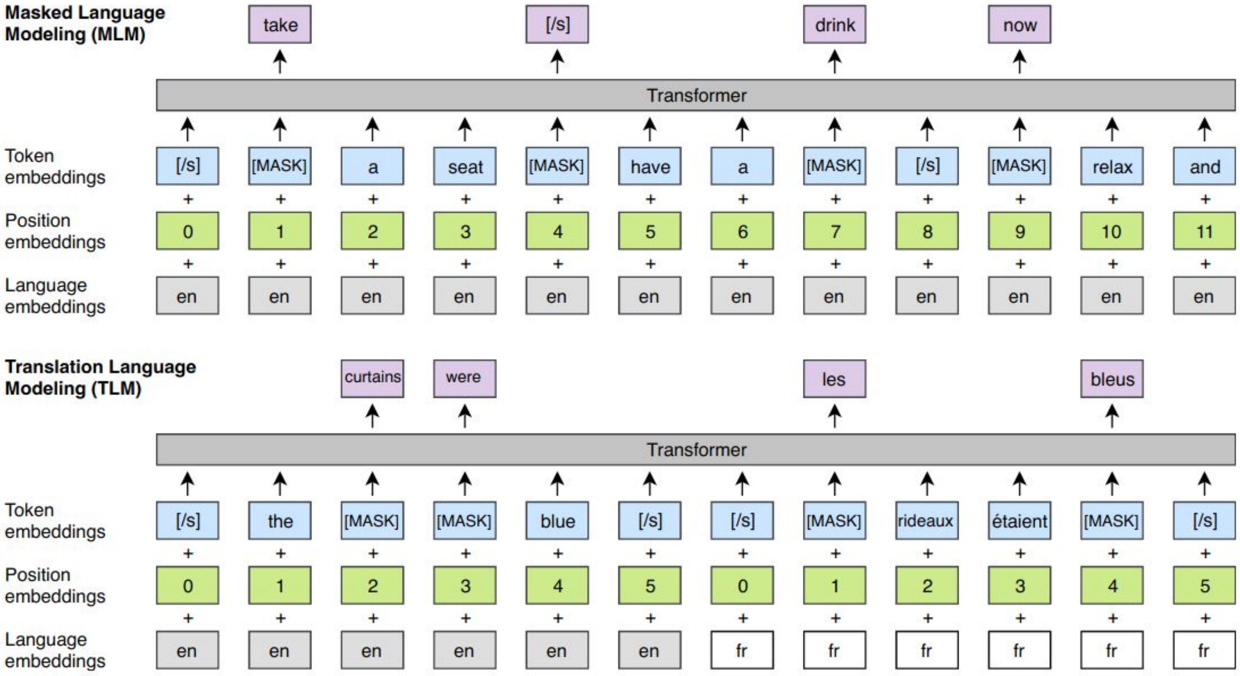


Figure 1: **Cross-lingual language model pretraining.** The MLM objective is similar to the one of Devlin et al. (2018), but with continuous streams of text as opposed to sentence pairs. The TLM objective extends MLM to pairs of parallel sentences. To predict a masked English word, the model can attend to both the English sentence and its French translation, and is encouraged to align English and French representations. Position embeddings of the target sentence are reset to facilitate the alignment.

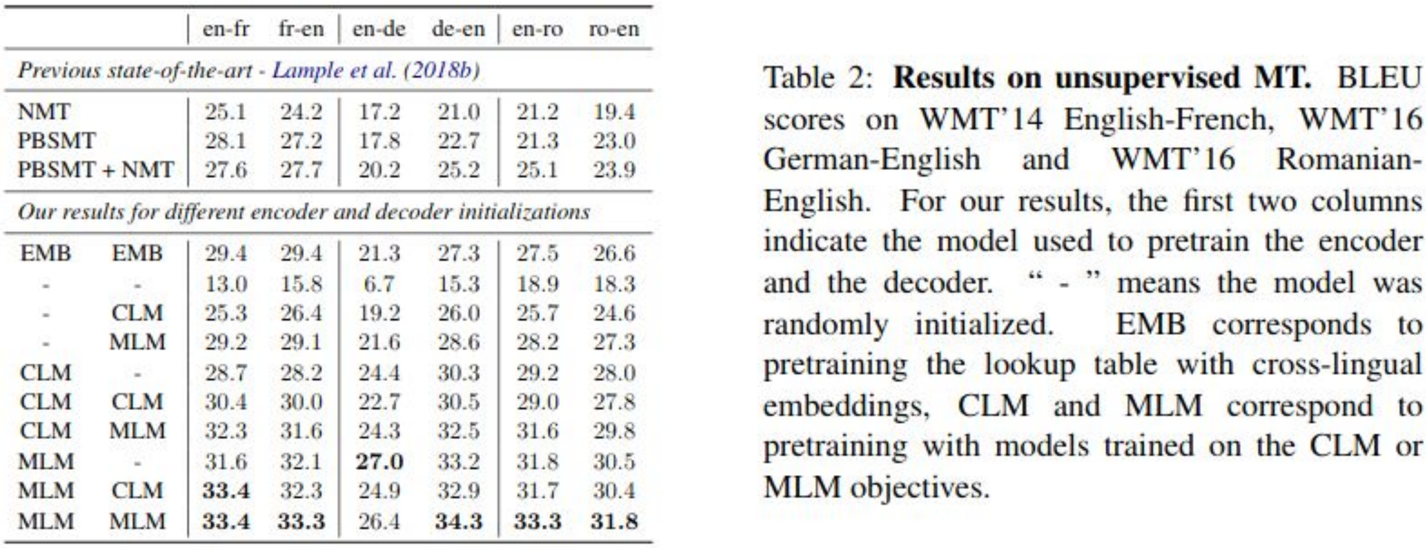
Notice: additional Language Embedding.

MLM is unsupervised, it does not require pairs of translated sentences in two languages.

TLM is supervised.

Unsupervised Machine Translation

Using models trained with MLM, they obtain a model that does "unsupervised" machine translation.



		en-fr	fr-en	en-de	de-en	en-ro	ro-en
Previous state-of-the-art - Lample et al. (2018b)							
NMT		25.1	24.2	17.2	21.0	21.2	19.4
PBSMT		28.1	27.2	17.8	22.7	21.3	23.0
PBSMT + NMT		27.6	27.7	20.2	25.2	25.1	23.9
Our results for different encoder and decoder initializations							
EMB	EMB	29.4	29.4	21.3	27.3	27.5	26.6
-	-	13.0	15.8	6.7	15.3	18.9	18.3
-	CLM	25.3	26.4	19.2	26.0	25.7	24.6
-	MLM	29.2	29.1	21.6	28.6	28.2	27.3
CLM	-	28.7	28.2	24.4	30.3	29.2	28.0
CLM	CLM	30.4	30.0	22.7	30.5	29.0	27.8
CLM	MLM	32.3	31.6	24.3	32.5	31.6	29.8
MLM	-	31.6	32.1	**27.0**	33.2	31.8	30.5
MLM	CLM	**33.4**	32.3	24.9	32.9	31.7	30.4
MLM	MLM	**33.4**	**33.3**	26.4	**34.3**	**33.3**	**31.8**

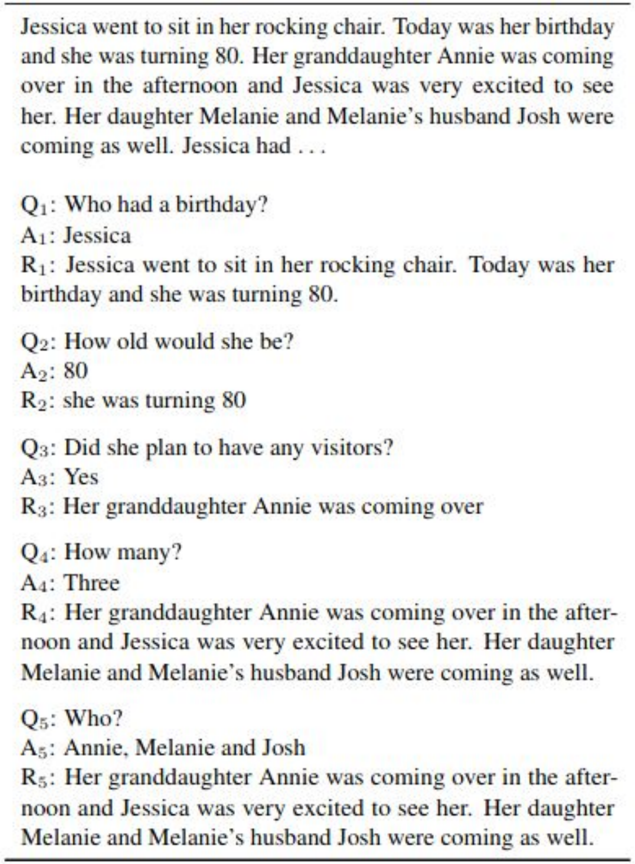
Table 2: **Results on unsupervised MT.** BLEU scores on WMT'14 English-French, WMT'16 German-English and WMT'16 Romanian-English. For our results, the first two columns indicate the model used to pretrain the encoder and the decoder. " - " means the model was randomly initialized. EMB corresponds to pretraining the lookup table with cross-lingual embeddings, CLM and MLM correspond to pretraining with models trained on the CLM or MLM objectives.

Conversational Q&A

https://stanfordnlp.github.io/coqa/

An conversation about a passage of text between an student and a teacher. The teacher must both generate an answer, and extract a relevant passage in the text.

Some challenges: coreference (to previous questions), reasoning, and abstractive generation.



Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well. Jessica had . . .

Q_1: Who had a birthday?
A_1: Jessica
R_1: Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.

Q_2: How old would she be?
A_2: 80
R_2: she was turning 80

Q_3: Did she plan to have any visitors?
A_3: Yes
R_3: Her granddaughter Annie was coming over

Q_4: How many?
A_4: Three
R_4: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

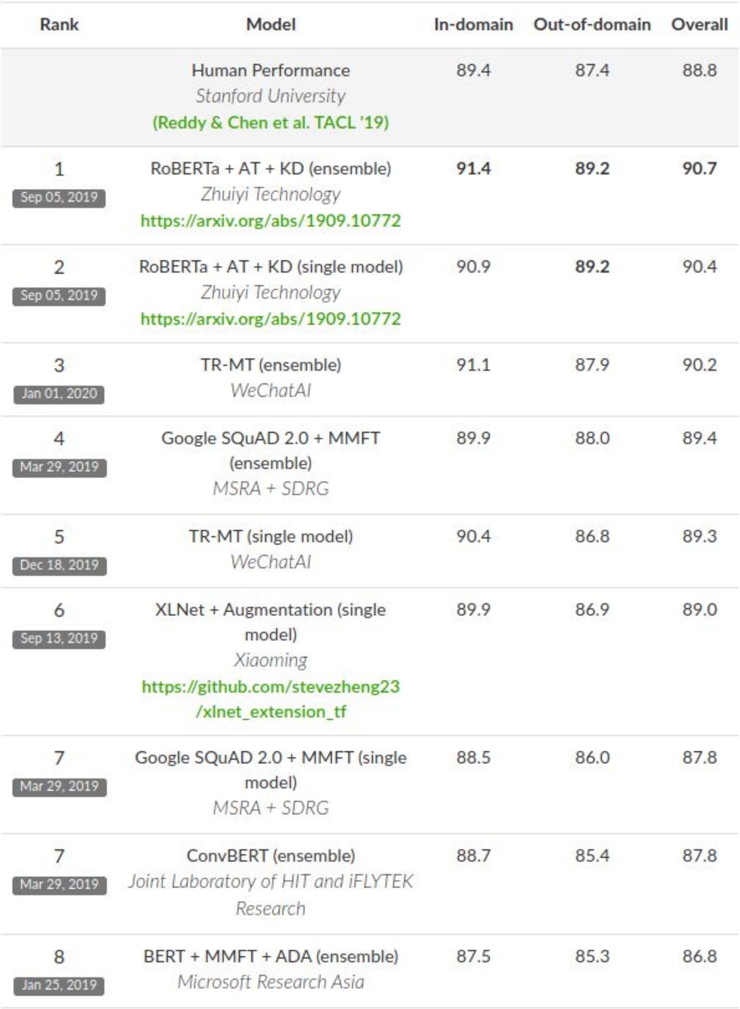
Q_5: Who?
A_5: Annie, Melanie and Josh
R_5: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Figure 1: A conversation from the CoQA dataset. Each turn contains a question (Q_i), an answer (A_i) and a rationale (R_i) that supports the answer.

CoQA Leaderboard

Notice: most models are BERT based (XLNet and RoBERTa are small variants).

Ensemble often gives a boost.



Rank	Model	In-domain	Out-of-domain	Overall
	Human Performance *Stanford University* (Reddy & Chen et al. TACL '19)	89.4	87.4	88.8
1 Sep 05, 2019	RoBERTa + AT + KD (ensemble) *Zhuiyi Technology* https://arxiv.org/abs/1909.10772	**91.4**	**89.2**	**90.7**
2 Sep 05, 2019	RoBERTa + AT + KD (single model) *Zhuiyi Technology* https://arxiv.org/abs/1909.10772	90.9	**89.2**	90.4
3 Jan 01, 2020	TR-MT (ensemble) *WeChatAI*	91.1	87.9	90.2
4 Mar 29, 2019	Google SQuAD 2.0 + MMFT (ensemble) *MSRA + SDRG*	89.9	88.0	89.4
5 Dec 18, 2019	TR-MT (single model) *WeChatAI*	90.4	86.8	89.3
6 Sep 13, 2019	XLNet + Augmentation (single model) *Xiaoming* https://github.com/stevezheng23/xlnet_extension_tf	89.9	86.9	89.0
7 Mar 29, 2019	Google SQuAD 2.0 + MMFT (single model) *MSRA + SDRG*	88.5	86.0	87.8
7 Mar 29, 2019	ConvBERT (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	88.7	85.4	87.8
8 Jan 25, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	87.5	85.3	86.8

Dialog

DSTC = Dialog System Technology Challenge:

- Sentence selection
- Sentence generation
- Audio-visual scene-aware dialog

Sample tasks for DSTC

ADVISOR | Hi! What can I help you with?

STUDENT | Hello! I'm trying to schedule classes for next semester. Can you help me?

STUDENT | Hardware has been an interest of mine.

STUDENT | But I don't want too hard of classes

ADVISOR | So are you interested in pursuing Electrical or Computer Engineering?

STUDENT | I'm undecided

STUDENT | I enjoy programming but enjoy hardware a little more.

ADVISOR | Computer Engineering consists of both programming and hardware.

ADVISOR | I think it will be a great fit for you.

STUDENT | Awesome, I think that's some good advice.

STUDENT | What classes should I take to become a Computer Engineer?

ADVISOR | You haven't taken EECS 203, 280, and 270, so it may be in your best interest to take one or two of those classes next semester

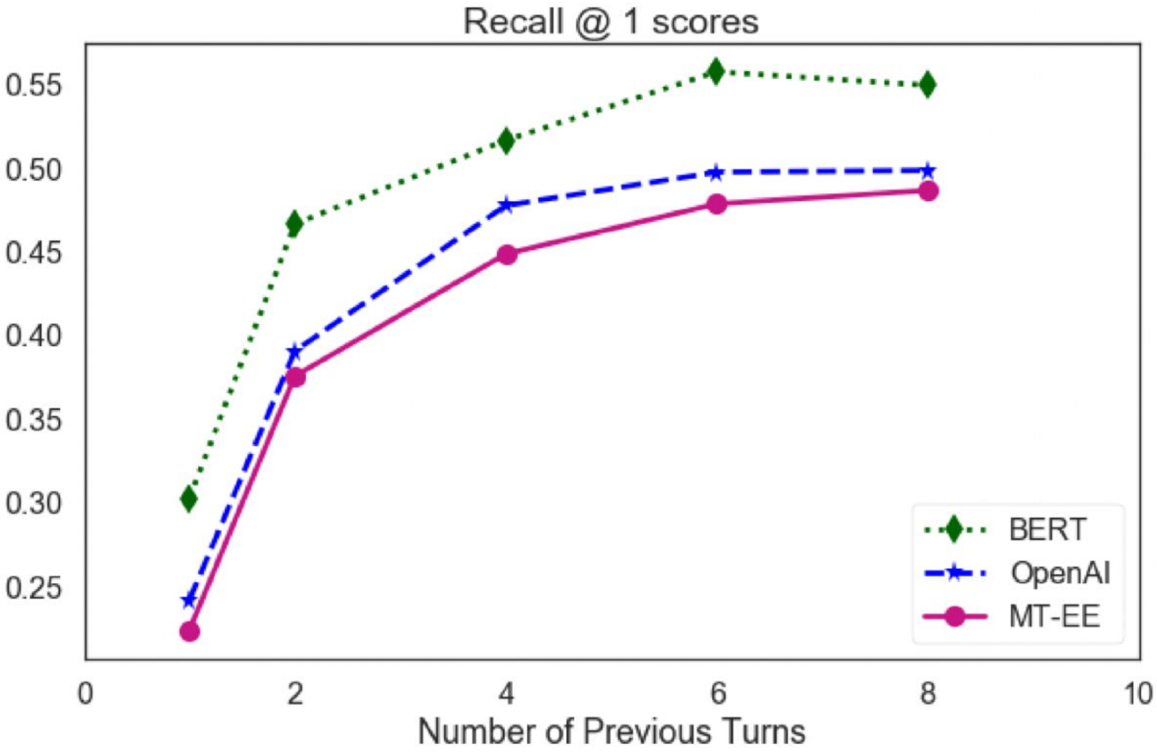
STUDENT | Ok. Which of those is in the morning. I like morning classes

Sample tasks for DSTC

```
[13:11] <user_1> anyone here know memcached?
[13:12] <user_1> trying to change the port it runs on
[13:12] <user_2> user_1: and ?
[13:13] <user_1> user_2: I'm not sure where to look
[13:13] <user_1> !
[13:13] <user_2> user_1: /etc/memcached.conf ?
[13:13] <user_1> haha
[13:13] <user_1> user_2: oh yes, it's much simpler than I thought
[13:13] <user_1> not sure why, I was trying to work through the init.d stuff
```

Should also use external reference information from unix man pages.

DSTC7 results



Takeaways

- Pre-training language models on very large corpora improves the state-of-the-art for multiple NLP tasks (ELMo, GPT, BERT).
- Transformer designs (GPT and BERT) have superseded RNN designs (ELMo).
- Single-task execution involves input encoding, a small amount of output hardware, and fine-tuning.
- BiDirectional encoder models (BERT) do better than generative models (GPT) at non-generation tasks, for comparable training data/model complexity.
- Generative models have training efficiency and scalability advantages that may make them ultimately more accurate. They can also solve entirely new kinds of task that involve text generation.