

# COLX-585: Trends in Computational Linguistics

**Muhammad Abdul-Mageed**

[muhmmad.mageed@ubc.ca](mailto:muhmmad.mageed@ubc.ca)

Natural Language Processing Lab

The University of British Columbia

# Generative Deep Learning: Language Models

# Table of Contents

1 Early Work

2 GPT-1

3 GPT-2

4 GPT-2 Samples

# RNN for Language Generation

## Generating Sequences With Recurrent Neural Networks

Alex Graves

Department of Computer Science

University of Toronto

[graves@cs.toronto.edu](mailto:graves@cs.toronto.edu)

### **Abstract**

This paper shows how Long Short-term Memory recurrent neural networks can be used to generate complex sequences with long-range structure, simply by predicting one data point at a time. The approach is demonstrated for text (where the data are discrete) and online handwriting (where the data are real-valued). It is then extended to handwriting synthesis by allowing the network to condition its predictions on a text sequence. The resulting system is able to generate highly realistic cursive handwriting in a wide variety of styles.

Figure: Graves (2013)

# RNN Architecture

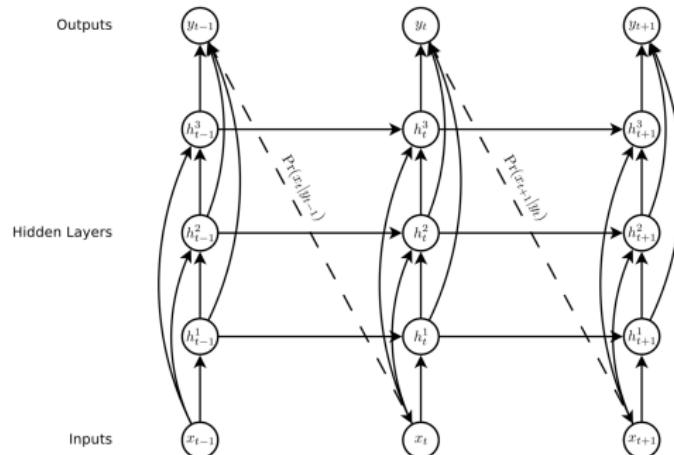


Figure 1: **Deep recurrent neural network prediction architecture.** The circles represent network layers, the solid lines represent weighted connections and the dashed lines represent predictions.

**Note:** the ‘skip connections’ from the inputs to all hidden layers, and from all hidden layers to the outputs. These mitigate ‘vanishing gradients’.

## Why Char Level?

- ① Modeling words would run into data sparsity
- ② For softmax-based models, high computational cost for evaluating all exponentials during training
- ③ Predicting one char at a time, allows the network to invent novel 'words'

## Eval

- ① Bits-per-character (BPC): = avg value of  $\log_2 p(x_{i+1}|y_i)$  over the whole test set
- ② Perplexity: Two to the power of the average number of bits per word
  - For a test set with avg. number of 5.6 chars, perplexity  $\approx 2^{5.6BPC}$

# Sample Generated Wikipedia Article

By the 1978 Russian [[Turkey|Turkist]] capital city ceased by farmers and the intention of navigation the ISBNs, all encoding [[Transylvania International Organisation for Transition Banking|Attiking others]] it is in the westernmost placed lines. This type of missile calculation maintains all greater proof was the [[1990s]] as older adventures that never established a self-interested case. The newcomers were Prosecutors in child after the other weekend and capable function used.

Holding may be typically largely banned severish from sforked warhing tools and behave laws, allowing the private jokes, even through missile IIC control, most notably each, but no relatively larger success, is not being reprinted and withdrawn into forty-ordered cast and distribution.

Besides these markets (notably a son of humor).

Sometimes more or only lowed " to force a suit for <http://news.bbc.co.uk/1/sid9kcid/web/9960219.html>".  
&lt;blockquote&gt;

==The various disputes between Basic Mass and Council Conditioners - "Titanist" class streams and anarchism==

Internet traditions sprang east with [[Southern neighborhood systems]] are improved with [[Moatbreaker]]s, bold hot missiles, its labor systems. [[KCD]] numbered former ISBN/MAS/speaker attacks "M3 5"; which are saved as the ballistic misely known and most functional factories. Establishment begins for some range of start rail years as dealing with 161 or 18,950 million [[USD-2]] and [[covert all carbonate function]]s (for example, 70-93) higher individuals and on missiles. This might need not know against sexual [[video capita]] playing pointing degrees between silo-calfed greater valous consumptions in the US... header can be seen in [[collectivist]].

-- See also --

**Figure:** Note: Top part not shows, but it has Wikipedia-like markup.

# Sample Generated Handwriting

more of national temperament

**Figure:** Top line is real and the rest are generated

# Primed Samples

*Take the breath away when they are*

*when the network is primed  
and biased, it writes  
in a cleaned up version  
of the original style*

*She looked closely as she*

*when the network is primed  
and biased, it writes  
in a cleaned up version  
of the original style*

Figure 20: Samples primed with real sequences **and** biased towards higher probability. The priming sequences are at the top of the blocks. The probability bias was 1. None of the lines in the sampled text exist in the training set.

# Semi-Supervised Sequence Learning

---

## Semi-supervised Sequence Learning

---

**Andrew M. Dai**

Google Inc.

adai@google.com

**Quoc V. Le**

Google Inc.

qvl@google.com

### Abstract

We present two approaches to use unlabeled data to improve Sequence Learning with recurrent networks. The first approach is to predict what comes next in a sequence, which is a language model in NLP. The second approach is to use a sequence autoencoder, which reads the input sequence into a vector and predicts the input sequence again. These two algorithms can be used as a “pretraining” algorithm for a later supervised sequence learning algorithm. In other words, the parameters obtained from the pretraining step can then be used as a starting point for other supervised training models. In our experiments, we find that long short term memory recurrent networks after pretrained with the two approaches become more stable to train and generalize better. With pretraining, we were able to achieve strong performance in many classification tasks, such as text classification with IMDB, DBpedia or image recognition in CIFAR-10.

Figure: (2015)

# Main Ideas

## Two Models

- ① Sequence auto-encoder
- ② Language Model, with LSTM (LM-LSTM)

# Sequence Auto-Encoder

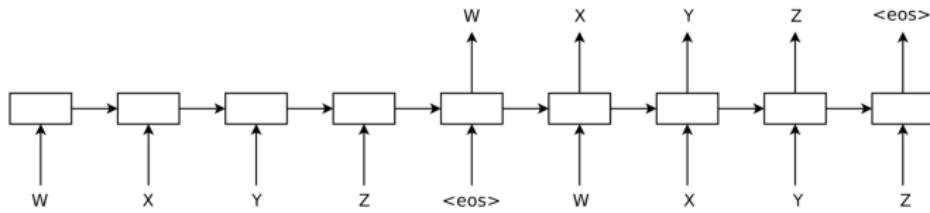


Figure 1: The sequence autoencoder for the sequence “WXYZ”. The sequence autoencoder uses a recurrent network to read the input sequence into the hidden state, which can then be used to reconstruct the original sequence.

## Sample Results: SAA

Table 1: A summary of the error rates of SA-LSTMs and previous best reported results.

<b>Dataset</b>	<b>SA-LSTM</b>	<b>Previous best result</b>
IMDB	7.24%	7.42%
Rotten Tomatoes	16.7%	18.5%
20 Newsgroups	15.6%	17.1%
DBpedia	1.19%	1.74%

## Sample Results: SAA LM

Table 6: Performance of models on the 20 newsgroups classification task.

<b>Model</b>	<b>Test error rate</b>
LSTM	18.0%
LM-LSTM	15.3%
LSTM with linear gain	71.6%
SA-LSTM	15.6%
Hybrid Class RBM [18]	23.8%
RBM-MLP [5]	20.5%
SVM + Bag-of-words [3]	17.1%
Naïve Bayes [3]	19.0%

## Improving Language Understanding by Generative Pre-Training

---

Alec Radford

OpenAI

alec@openai.com

Karthik Narasimhan

OpenAI

karthikn@openai.com

Tim Salimans

OpenAI

tim@openai.com

Ilya Sutskever

OpenAI

ilyasu@openai.com

### Abstract

Natural language understanding comprises a wide range of diverse tasks such as textual entailment, question answering, semantic similarity assessment, and document classification. Although large unlabeled text corpora are abundant, labeled data for learning these specific tasks is scarce, making it challenging for discriminatively trained models to perform adequately. We demonstrate that large gains on these tasks can be realized by *generative pre-training* of a language model on a diverse corpus of unlabeled text, followed by *discriminative fine-tuning* on each specific task. In contrast to previous approaches, we make use of task-aware input transformations during fine-tuning to achieve effective transfer while requiring minimal changes to the model architecture. We demonstrate the effectiveness of our approach on a wide range of benchmarks for natural language understanding. Our general task-agnostic model outperforms discriminatively trained models that use architectures specifically crafted for each task, significantly improving upon the state of the art in 9 out of the 12 tasks studied. For instance, we achieve absolute improvements of 8.9% on commonsense reasoning (Stories Cloze Test), 5.7% on question answering (RACE), and 1.5% on textual entailment (MultiNLI).

Figure: (2018)

# Key Points

## Two Models

- ① Uses Transformer:** capture longer-range linguistic structure
- ② Application on a wide range of tasks:** natural language inference, paraphrase detection and story completion
- ③ Use of large dataset**

## Recall: Language Modeling Objective

Given a sequence of tokens  $\mathcal{U} = \{u_1, \dots, u_n\}$ , a standard language modeling objective maximizes the following log likelihood:

### 1: LM Objective

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \dots, u_{i-1}; \theta)$$

- where  $k$  is the size of the context window, and  $P$  is modeled with a neural network (here a **Transformer decoder**) with parameters  $\theta$

# LM Unsupervised Pre-Training With Transformer Decoder

Apply multi-head attention on input context tokens, followed by a position-wise feedforward network to produce an output distribution over tokens:

## 2: LM With Transformer

$$h_0 = UW_e + W_p$$

$$h_i = \text{transformer\_block}(h_{i-1}) \forall i \in [1, n]$$

$$P(u) = \text{softmax}(h_n W_e^T)$$

- $U = (u_k, \dots, u_i)$ : context vector of tokens
- $n$ : number of layers
- $W_e$ : Token embedding matrix;  $W_p$ : Position embedding matrix

# Supervised Fine-Tuning

## Two Models

Assuming a **dataset  $C$** , with **input tokens  $x^1, \dots, x^m$** , along with a **label  $y$** :

- ① **Inputs are passed through the pre-trained model to acquire final transformer's block activation  $h_I^m$ ,**
- ② which is fed to an added **linear output layer** with parameters  $W_y$  **to predict  $y$**

## 3: Supervised Fine-Tuning

$$P(y|x^1, \dots, x^m) = \text{softmax}(h_I^m W_y)$$

This gives the following objective:

$$L_2(C) = \sum_{(x,y)} \log P(y|x^1, \dots, x^m)$$

# Language Modeling Auxiliary Objective

## 4: Supervised Fine-Tuning

Adding LM task objective:

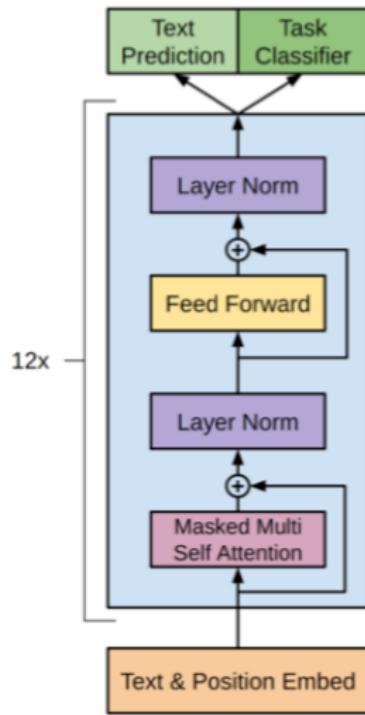
$$L_3(C) = L_2(C) + \lambda L_1(C)$$

### Recall

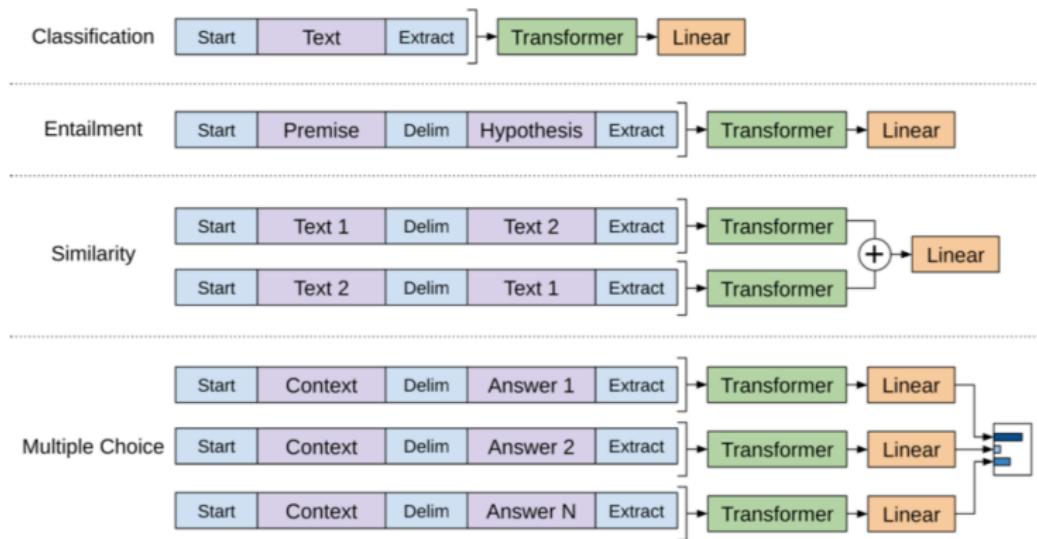
Assuming a dataset  $C$ , with input tokens  $x^1, \dots, x^m$ , along with a label  $y$ :

- $L_1$  is a language modeling objective
- $L_2$  itself is optimized with a Transformer trained for language modeling

# Overall architecture

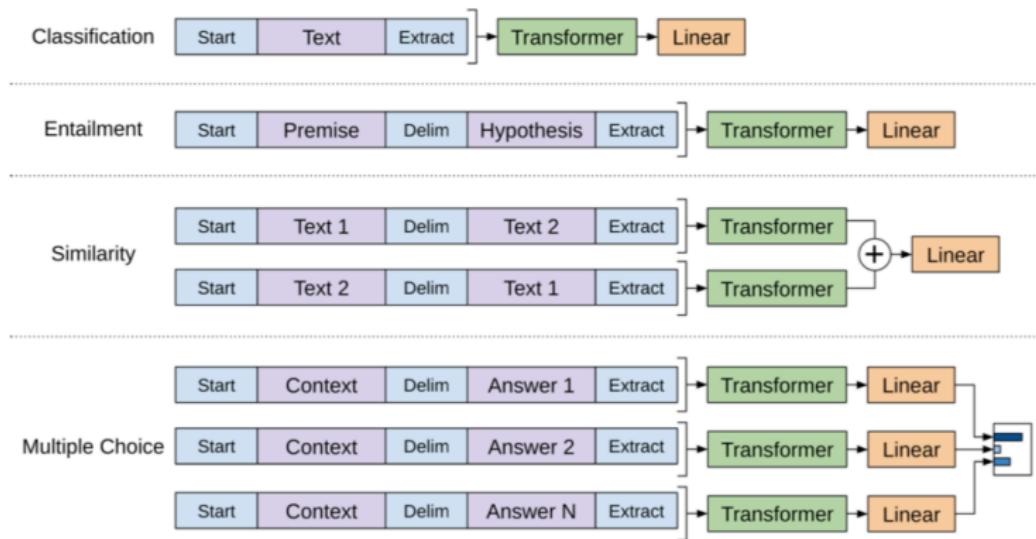


# Input Transformations for Fine-Tuning



**Figure:** Authors convert all structured inputs into token sequences to be processed by the pre-trained model, followed by a linear+softmax layer

# Input Transformations for Fine-Tuning



**Figure:** Authors convert all structured inputs into token sequences to be processed by the pre-trained model, followed by a linear+softmax layer

## FYI: Input Transformations Details

**Textual entailment** For entailment tasks, we concatenate the premise  $p$  and hypothesis  $h$  token sequences, with a delimiter token (\$) in between.

**Similarity** For similarity tasks, there is no inherent ordering of the two sentences being compared. To reflect this, we modify the input sequence to contain both possible sentence orderings (with a delimiter in between) and process each independently to produce two sequence representations  $h_l^m$  which are added element-wise before being fed into the linear output layer.

**Question Answering and Commonsense Reasoning** For these tasks, we are given a context document  $z$ , a question  $q$ , and a set of possible answers  $\{a_k\}$ . We concatenate the document context and question with each possible answer, adding a delimiter token in between to get  $[z; q; \$; a_k]$ . Each of these sequences are processed independently with our model and then normalized via a softmax layer to produce an output distribution over possible answers.

# Model Details

## Dataset

- ① **BooksCorpus dataset:** unpublished 7,000 books from a variety of genres including Adventure, Fantasy, and Romance
- ② **Long stretches of contiguous text:** allows the generative model to learn to condition on long-range information

**Model specifications** Our model largely follows the original transformer work [62]. We trained a 12-layer decoder-only transformer with masked self-attention heads (768 dimensional states and 12 attention heads). For the position-wise feed-forward networks, we used 3072 dimensional inner states. We used the Adam optimization scheme [27] with a max learning rate of 2.5e-4. The learning rate was increased linearly from zero over the first 2000 updates and annealed to 0 using a cosine schedule. We train for 100 epochs on minibatches of 64 randomly sampled, contiguous sequences of 512 tokens. Since layernorm [2] is used extensively throughout the model, a simple weight initialization of  $N(0, 0.02)$  was sufficient. We used a bytepair encoding (BPE) vocabulary with 40,000 merges [53] and residual, embedding, and attention dropouts with a rate of 0.1 for regularization. We also employed a modified version of L2 regularization proposed in [37], with  $w = 0.01$  on all non bias or gain weights. For the activation function, we used the Gaussian Error Linear Unit (GELU) [18]. We used learned position embeddings instead of the sinusoidal version proposed in the original work. We use the *ftfy* library<sup>[2]</sup> to clean the raw text in BooksCorpus, standardize some punctuation and whitespace, and use the *spaCy* tokenizer.<sup>[3]</sup>

# NLI Results

Table 2: Experimental results on natural language inference tasks, comparing our model with current state-of-the-art methods. 5x indicates an ensemble of 5 models. All datasets use accuracy as the evaluation metric.

Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	<u>89.3</u>	-	-	-
CAFE [58] (5x)	80.2	79.0	<u>89.3</u>	-	-	-
Stochastic Answer Network [35] (3x)	<u>80.6</u>	<u>80.1</u>	-	-	-	-
CAFE [58]	78.7	77.9	88.5	<u>83.3</u>		
GenSen [64]	71.4	71.3	-	-	<u>82.3</u>	59.2
Multi-task BiLSTM + Attn [64]	72.2	72.1	-	-	82.1	<b>61.7</b>
Finetuned Transformer LM (ours)	<b>82.1</b>	<b>81.4</b>	<b>89.9</b>	<b>88.3</b>	<b>88.1</b>	56.0

# Q&A and Commonsense Reasoning Results

Table 3: Results on question answering and commonsense reasoning, comparing our model with current state-of-the-art methods.. 9x means an ensemble of 9 models.

Method	Story Cloze	RACE-m	RACE-h	RACE
val-LS-skip [55]	76.5	-	-	-
Hidden Coherence Model [7]	<u>77.6</u>	-	-	-
Dynamic Fusion Net [67] (9x)	-	55.6	49.4	51.2
BiAttention MRU [59] (9x)	-	<u>60.2</u>	<u>50.3</u>	<u>53.3</u>
Finetuned Transformer LM (ours)	<b>86.5</b>	<b>62.9</b>	<b>57.4</b>	<b>59.0</b>

---

## Language Models are Unsupervised Multitask Learners

---

Alec Radford <sup>\*†</sup> Jeffrey Wu <sup>\*†</sup> Rewon Child <sup>†</sup> David Luan <sup>†</sup> Dario Amodei <sup>\*\*†</sup> Ilya Sutskever <sup>\*\*†</sup>

### Abstract

Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset - matching or exceeding the performance of 3 out of 4 baseline systems without using the 127,000+ training examples. The capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linear fashion across tasks. Our largest model, GPT-2, is a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting but still underfits WebText. Samples from the model reflect these improvements and contain coherent paragraphs of text. These findings suggest a promising path towards building language processing systems which learn to perform tasks from their naturally occurring demonstrations.

competent generalists. We would like to move towards more general systems which can perform many tasks – eventually without the need to manually create and label a training dataset for each one.

The dominant approach to creating ML systems is to collect a dataset of training examples demonstrating correct behavior for a desired task, train a system to imitate these behaviors, and then test its performance on independent and identically distributed (IID) held-out examples. This has served well to make progress on narrow experts. But the often erratic behavior of captioning models (Lake et al., 2017), reading comprehension systems (Jia & Liang, 2017), and image classifiers (Alcorn et al., 2018) on the diversity and variety of possible inputs highlights some of the shortcomings of this approach.

Our suspicion is that the prevalence of single task training on single domain datasets is a major contributor to the lack of generalization observed in current systems. Progress towards robust systems with current architectures is likely to require training and measuring performance on a wide range of domains and tasks. Recently, several benchmarks have been proposed such as GLUE (Wang et al., 2018) and decaNLP (McCann et al., 2018) to begin studying this.

Multitask learning (Caruana, 1997) is a promising framework for improving general performance. However, multitask training in NLP is still nascent. Recent work reports modest performance improvements (Yogatama et al.,

# Key Points

## Zero-Shot Learning & Input Task Transformation/Re-Writing

- ① Language models can perform down-stream tasks in zero-shot settings
- ② Language models as **multi-task learners**
- ③ Perhaps better: Powerful language models learn **universal language representations**
- ④ Most NLP tasks (all?) can be cast as a **question answering problem**
- ⑤ **Translation:** **can be re-written as** → (translate to french, english text, french text)
- ⑥ **Reading comprehension:** → (answer the question, document, question, answer)
- ⑦ For applications of this re-writing approach, see Kumar et al. (2016); McCann et al. (2018)

# NLP Tasks as QA Problems

---

## Ask Me Anything: Dynamic Memory Networks for Natural Language Processing

---

Ankit Kumar\*, Ozan Irsoy\*, Peter Ondruska\*,  
Mohit Iyyer\*, James Bradbury, Ishaan Gulrajani\*,  
Victor Zhong\*, Romain Paulus, Richard Socher {JAMES.BRADBURY,RPAULUS,RSOCHER}@SALESFORCE.COM  
Salesforce Inc., CA USA

### Abstract

Most tasks in natural language processing can be cast into question answering (QA) problems over language input. We introduce the dynamic memory network (DMN), a neural network architecture which processes input sequences and questions, forms episodic memories, and generates relevant answers. Questions trigger an iterative attention process which allows the model to condition its attention on the inputs and the result of previous iterations. These results are then reasoned over in a hierarchical recurrent sequence model to generate answers. The DMN can be trained end-to-end and obtains state-of-the-art results on several types of tasks and datasets: question answering (Facebook's bAbI dataset), text classification for sentiment analysis (Stanford Sentiment Treebank) and sequence modeling for part-of-speech tagging (WSJ-PTB). The training for these different tasks relies exclusively on trained word vector representations and input-question-answer triplets.

- I: Jane went to the hallway.  
I: Mary walked to the bathroom.  
I: Sandra went to the garden.  
I: Daniel went back to the garden.  
I: Sandra took the milk there.  
Q: Where is the milk?  
A: garden  
I: It started boring, but then it got interesting.  
Q: What's the sentiment?  
A: positive  
Q: POS tags?  
A: PRP VBD JJ , CC RB PRP VBD JJ .

Figure 1. Example inputs and questions, together with answers generated by a dynamic memory network trained on the corresponding task. In sequence modeling tasks, an answer mechanism is triggered at each input word instead of only at the end.

tion problems like sentiment analysis (Socher et al., 2013) (*What is the sentiment?*); even multi-sentence joint classification problems like coreference resolution (*Who does "their" refer to?*).

Figure: [ICML. 2016]

# Training Dataset

## WebText

- ① Scraped **outbound links from Reddit** which received at least 3 karma
- ② **WebText**, contains the text subset of these **45 million links**
- ③ Does not include links created after Dec 2017
- ④ After de-duplication and some heuristic based cleaning contains slightly over **8 million documents for a total of 40 GB** of text.
- ⑤ **Removed all Wikipedia documents** from WebText since it is a common data source for other datasets

# Translation Data in WebText

---

"I'm not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile** [I'm not a fool].

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: "**Mentez mentez, il en restera toujours quelque chose**," which translates as, "Lie lie and something will always remain."

"I hate the word '**perfume**,'" Burr says. 'It's somewhat better in French: '**parfum**'.

If listened carefully at 29:55, a conversation can be heard between two guys in French: "**-Comment on fait pour aller de l'autre côté? -Quel autre côté?**", which means "- How do you get to the other side? - What side?".

If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

**"Brevet Sans Garantie Du Gouvernement"**, translated to English: **"Patented without government warranty"**.

---

*Table 1.* Examples of naturally occurring demonstrations of English to French and French to English translation found throughout the WebText training set.

# Model Details

## Dataset

① Vocabulary: BPE,  $V = 50,257$

② Authors train 4 models:

Parameters	Layers	$d_{model}$
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600

Table 2. Architecture hyperparameters for the 4 model sizes.

# Sample Results

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

# Sample Answers | Picks Name From Doc...

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calendar is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%
What us state forms the western boundary of montana?	Montana	✗	52.3%
Who plays ser davos in game of thrones?	Peter Dinklage	✗	52.1%
Who appoints the chair of the federal reserve system?	Janet Yellen	✗	51.5%
State the process that divides one nucleus into two genetically identical nuclei?	mitosis	✓	50.7%
Who won the most mvp awards in the nba?	Michael Jordan	✗	50.2%
What river is associated with the city of rome?	the Tiber	✓	48.6%
Who is the first president to be impeached?	Andrew Johnson	✓	48.3%
Who is the head of the department of homeland security 2017?	John Kelly	✓	47.0%
What is the name given to the common currency to the european union?	Euro	✓	46.8%
What was the emperor name in star wars?	Palpatine	✓	46.5%
Do you have to have a gun permit to shoot at a range?	No	✓	46.4%
Who proposed evolution in 1859 as the basis of biological development?	Charles Darwin	✓	45.7%
Nuclear power plant that blew up in russia?	Chernobyl	✓	45.7%
Who played john connor in the original terminator?	Arnold Schwarzenegger	✗	45.2%

Table 5. The 30 most confident answers generated by GPT-2 on the development set of Natural Questions sorted by their probability according to GPT-2. None of these questions appear in WebText according to the procedure described in Section 4.



## MT Results

We test whether GPT-2 has begun to learn how to translate from one language to another. In order to help it infer that this is the desired task, we condition the language model on a context of example pairs of the format `english sentence = french sentence` and then after a final prompt of `english sentence =` we sample from the model with greedy decoding and use the first generated sentence as the translation. On the WMT-14 English-French test set, GPT-2 gets 5 BLEU, which is slightly worse than a word-by-word substitution with a bilingual lexicon inferred in previous work on unsupervised word translation

## MT Results *Contd.*

(Conneau et al., 2017b). On the WMT-14 French-English test set, GPT-2 is able to leverage its very strong English language model to perform significantly better, achieving 11.5 BLEU. This outperforms several unsupervised machine translation baselines from (Artetxe et al., 2017) and (Lample et al., 2017) but is still much worse than the 33.5 BLEU of the current best unsupervised machine translation approach (Artetxe et al., 2019). Performance on this task was surprising to us, since we deliberately removed non-English webpages from WebText as a filtering step. In order to confirm this, we ran a byte-level language detector<sup>2</sup> on WebText which detected only 10MB of data in the French language which is approximately 500x smaller than the monolingual French corpus common in prior unsupervised machine translation research.

# GPT-2 Samples

SYSTEM PROMPT (HUMAN-WRITTEN)	<p><i>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.</i></p>
MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)	<p>The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.</p> <p>Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.</p> <p>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.</p> <p>Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.</p>

Figure: Try it out at <https://talktotransformer.com>.

# GPT-2 Sample *Contd.*

Custom prompt

After about 5 years of absence, it turns out that I can still drive without a GPS in Bloomington

GENERATE ANOTHER

## Completion

**After about 5 years of absence, it turns out that I can still drive without a GPS in Bloomington.** I just have to tell the computer what to do. That's right, I'm using Android, so I actually have to push the button to start the car and it works. And then I just tell it to take me where I want to go with my GPS.

Not too long ago I was driving an F150 pickup from Madison to Indiana, so I decided that with the GPS in, it would be fun to go to my friend's house, check the weather, grab a drink and then I could go back to my own place. I remember that the GPS was on in his truck, but I was driving a long way too slowly.

After much research, it was determined that there were a few things that would help for that type of situation. The first being I had to be able to tell the computer what to do. A simple push of the button would allow the computer to start the car. The second was some quick thinking on my part. I was sitting in the driver seat of an iPhone and decided that I would push the button.

# GPT-2 Samples *Contd.*

Photography is the art of capturing light with a camera.

**GENERATE ANOTHER**

## Completion

**Photography is the art of capturing light with a camera.** Photography is used as a means to a desired end, for two main reasons.

The first is that people and places in photos provide us with immediate and tangible evidence. A close-up of the perfect place, the kind that makes a painting seem twice as powerful, or a long-exposure photograph that dramatizes a moment in time is far more powerful than a stationary snapshot or still image. Photographs provide a more visceral experience with their objects of the moment, the impressions, emotions, and feeling that our eyes take on while looking at the subject of the photo.

We can say that photography is a medium of perception. To do so, we must look at a photograph and

Photography is the art of capturing light with a camera?!!

[GENERATE ANOTHER](#)

## Completion

**Photography is the art of capturing light with a camera?!!** This is a photo taken by David Jenning at Honolulu a few days before he died in August 2011. "I was wondering if you would mind answering this question in relation to your concepts of Photography" David is a great student of the art of photography. His last two photography books are amazing . He did a lot of creating in the art and lighting of his photography. A lot of what you see here was shot with an EOS T3. Camera accessories? A lot of his work uses a heavy dose of lighting to make his images pop. [He was very sweet and gentle even when falling over last year, yet he had a joyful personality]. I bet he was a great photographer himself, and

## GPT-2 Samples *Contd.*

"Photography is the art of capturing light with a camera", said the detective. He then jumped on his jet.

[GENERATE ANOTHER](#)

### Completion

**"Photography is the art of capturing light with a camera", said the detective.**  
**He then jumped on his jet.** On its side, he had a book in which he was reading. When he was walking with that young detective, a book fell off his face. This is a thoughtless action that is harmful.

"No, I mean that...."

He looked at the book again. There was a section on suicide. It read. "The only way to keep yourself from hurting yourself is to commit suicide. Even if you won't be able to watch the great flame... suicide that light. The person who can force him to pass the burning intense flames and brave the burning burning flames will be the man who become the first man to

## GPT-2 Samples *Contd.*

"Photography is the art of capturing light with a camera", said the detective. He then jumped on his jet. His name was

[GENERATE ANOTHER](#)

### Completion

**"Photography is the art of capturing light with a camera", said the detective.  
He then jumped on his jet. His name was** Wally Sims.

Then I took my last photo. I'll never forget the following day's paper. It read:  
"Great Journey Begins With Eruption of Lake Burlingame".

Ladies and gentlemen, I present to you, of course, the Bruce Wright Archibald  
Photographer of the Year Award.

## GPT-2 Samples *Contd.*

The cave was so dark, so strange, and my blood

**GENERATE ANOTHER**

### Completion

**The cave was so dark, so strange, and my blood** slowly took on a haunting green hue. The tree branch and tent made a clattering noise, and the shrubs started shaking. I noticed all around me a strange landscape of nothingness, and I heard my own low screams, but I was not sure if they were real or I was simply imagining them. I did not know what happened. I wanted to scream too. Then I felt something heavy on my back, and the blackness of the cave merged with my vision again. What seemed like a white dot floated in the midst of the blackness. I tried to take the dot in my sight, but in my confusion I screamed. In my confusion, I hadn't realized I was on the