Unsupervised Natural Question Answering with a Small Model

Martin Andrews
Red Dragon AI
Singapore
martin@reddragon.ai

Sam Witteveen
Red Dragon AI
Singapore
sam@reddragon.ai

Abstract

The recent demonstration of the power of huge language models such as GPT-2 to memorise the answers to factoid questions raises questions about the extent to which knowledge is being embedded directly within these large models. This short paper describes an architecture through which much smaller models can also answer such questions - by making use of 'raw' external knowledge. The contribution of this work is that the methods presented here rely on unsupervised learning techniques, complementing the unsupervised training of the Language Model. The goal of this line of research is to be able to add knowledge explicitly, without extensive training.

1 Introduction

The field of question answering has been dominated by supervised methods for competitive tasks such as the Stanford question answering dataset (SQuAD) (Rajpurkar et al., 2016). However, as discussed in Yogatama et al. (2019), some of these datasets are becoming over-optimised for, making the architectures less generally applicable.

At the other extreme, the ability of the GPT-2 (Radford et al., 2019) model to answer factoid questions, based purely on unsupervised training directed at improving its Language Model (LM) performance, was striking. But further reflection highlights the following issues :

- Questions correctly (and confidently) answered were a small fraction ($\sim$1%) of the questions asked
- Huge model size and long training periods were required before such behaviour was manifested
- This does not appear to be a practical approach to adsorbing an extensive knowledge-base

This work describes early work in aiding generalised models such as GPT-2 to answer questions, without having to embed facts directly in the model's weights. The overall direction of work is towards encouraging such generalised models to make use of external datasources (and other resources) without having to internalise all the data in models of exponentially increasing size (e.g. GPT-2-1.5B is more than 10x the size of GPT-2-117M).

2 Natural Questions Dataset

The Natural Questions (NQ) dataset (Kwiatkowski et al., 2019) is a question answering dataset containing 307,373 training examples, 7,830 development examples, and 7,842 test examples. Each example is comprised of a google.com query and a corresponding Wikipedia page. Each Wikipedia page has a passage (or long answer) annotated on the page that answers the question and one or more short spans from the annotated passage containing the actual answer. The long and the short answer annotations can however be empty. If they are both empty, then there is no answer on the page at all. If the long answer annotation is non-empty, but the short answer annotation is empty, then the annotated passage answers the question but no explicit short answer could be found. Finally, 1% of the documents have a passage annotated with a short answer that is 'yes' or 'no', instead of a list of short spans.

As reported in Radford et al. (2019), GPT-2-1.5B answers 4.1% of NQ questions correctly when evaluated by the exact match metric commonly used on reading comprehension datasets like SQuAD. In contrast, the smallest GPT-2-117M model (used as the basis for the model proposed in this work) is reported as not being capa-

ble of exceeding the 1.0% accuracy of the simple baseline which returns the most common answer for each question type (who, what, where, etc...). The fact that GPT-2-1.5B answered 5.3 times more questions correctly suggests that model capacity has been a major factor in the poor performance of neural systems on this kind of task as of yet.

3 Model Architecture

The model proposed here is built from several components which include (a) 876k Wikipedia sentences, addressible via embeddings; (b) a pre-trained GPT-2-117M language model which was noted to be incapable of answering questions successfully in Radford et al. (2019); and (c) a scheme for incorporating 'sentence hints' into the language generation context.

3.1 Embeddings for Sentence Lookup

Three different embedding methods were used :

(i) pre-trained BERT-regular (Devlin et al., 2018), using the the `bert-as-service` Python tool[1]. For a given input sentence this returns a 768-d embedding, calculated as the GlobalAveragePooling of the top-but-one layer of the pretrained BERT model;

(ii) Smooth Inverse Frequency (SIF) (Arora et al., 2017) embeddings, calculated by inverse-frequency weighting the BPE embeddings (from the GPE-2-117M model being used for the text generation task) followed by removal of the first PCA component; and

(iii) Universal Sentence Encoder (Cer et al., 2018), the training details not clear in the paper, but USE is not a purely unsupervised model : "We augment unsupervised learning with training on supervised data from the Stanford Natural Language Inference (SNLI) corpus" (Bowman et al., 2015).

Methods (i) and (ii) were not fine-tuned on the question answering task (since this would violate the spirit of this unsupervised-only system), whereas method (iii) was included to judge the benefits of adding some supervised training to the embedding stage.

3.2 Embeddings for Questions

In order that facts might be supplied by external text, embeddings $e(s_n)$ were produced for each sentence s_n of the $N(= 876,645)$ wikitext sentences, and also $e(q_j)$ was calculated for each q_j of the J questions.

The search term was calculated by adding a 'question to sentence' vector, set to the mean difference between the embeddings for question phrases and those of wikitext sentences to the original question q_j :

$$search_j = e(q_j) + \frac{1}{N}\sum e(s_\cdot) - \frac{1}{J}\sum e(q_\cdot)$$

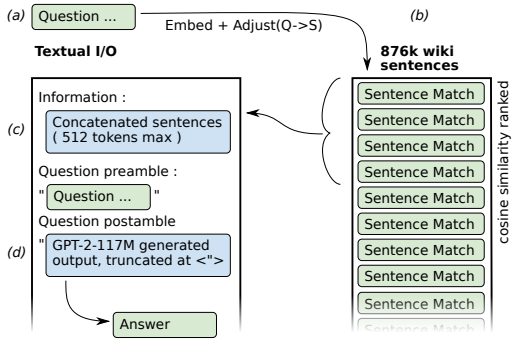

Figure 1: Proposed information flow : (a) Initial question; (b) Wiki sentence ranking; (c) hinting in preamble; (d) GPT2 output.

3.3 Knowledge Look-up

In order to aid the LM in retrieving factoid answers, 'hint sentences' sufficient to fill half of the LM context window were retrieved from the list of the N wikitext sentences, using a cosine distance ranking of the s_n vs $search_j$

3.4 LM Context Seeding

In order to obtain the results in Radford et al. (2019) for the NQ task, their GPT-2-1.5B model context was seeded with example question/answer pairs which helped the model infer the short answer style of the dataset.

Rather than expect the smaller GPT model to extrapolate from the Q & A format, both the 'hint sentences' and the question q_i were incorporated into the context seen by the model directly:

> Information :
>
> `HintSentence[ ]` *or* `None`
>
> The best short answer to "q_i?" from the information above is " ...

[1] https://bert-as-service.readthedocs.io/

Table 1: Sample question answers with filter examples, and examples of answers where pure SQuAD accuracy did not make sense when the base data included far more information than the original (single) wiki article targetted by the Natural Questions dataset.

Question	Target	GPT-2-117M	Reject reason
Who is the richest club in the championship?	'Aston Villa', 'Manchester City'	The richest club in the championship	SMART ALEC
Are all firestone tires made in the usa?	'NO'	No	Y/N QUESTION
What is the name of manchester united stadium?	'Old Trafford'	Manchester United	WITHIN QUESTION
Who cracked the enigma code in world war 2?	'Turing'	Alan Turing	N/A : ACCEPTED
How many inches is the iphone 5s screen?	'4 - inch screen size', '4 in', '4 in (10 cm)'	4 inches	N/A : ACCEPTED

The GPT-2-117M output is then recorded up until the closing double-quote (closing quotes appears to be strongly favoured by the LM).

3.5 Sampling from the Language Model

A number of approaches to sampling from the model were tried (including Beam search, which performed poorly), and the following were found to work satisfactorially :

1. SoftMax temperature was kept at 1.0 (i.e. as trained);

2. Nucleus Sampling (Holtzman et al., 2019) was used, with only tokens that cover the first 90% of probability space being considered as choices at each step. This appears to give a good mix of diversity without 'going off the rails' - which is desirable for human-like communication (Grice, 1975);

3. A probability bias term (Murray and Chiang, 2018) was added to the log-probabilities of each sequence, whereby each token was 'awarded' a bonus of α, which was found empirically to create a more balanced spread of long and short outputs;

4. After a sorted list of 100 different sequences was created, this was further filtered (as illustrated in Table 1) to reject answers that were very unlikely to be correct:

    - answers that simply repeat the question (determined as whether the answer's bigram Jaccard similarity with the question exceeds 0.5);
    - answers that are contained within the question verbatim;
    - answers such as 'yes/no', 'i don't know', 'none', 'no one', 'it depends' - which may have been safe choices, but could not score positively on the filtered list of questions.

Further details can be found in the Supplimental Materials.

4 Experiments

The model architecture was applied to the NQ task, and results are reported for performance on the validation set (the training set was unused). Only questions that were (a) not Yes/No; and (b) had a 'short answer' were considered, resulting in 3975 triples of {question, wikitext, answer list}.

The list of 'hint sentence' candidates was set to be the aggregate of all the sentences across the 3975 wikitext pages, totalling ~876k sentences. Importantly, the hint sentence choices weren't restricted to the wikitext corresponding to the specific question - which makes the task significantly more difficult that the BERT baseline for Natural Questions task (Alberti et al., 2019), which works on an article-by-article basis.

In the results reported, to reduce noise, the 'Yes/No' questions were removed from consideration (since scoring positively on these examples may the result of a coin-flip).

5 Results

This work is in its early stages, and the results obtained so far are encouraging, despite being low in number.

For the 3975 useful NQ development set questions, we found that the poor results of using GPT-2-117M unaided reported in Radford et al. (2019) were born out.

However, when using each question to select 'hint sentences' from the whole list of 876k wikitext sentences, the GPT-2-117M was able to make use of the extra information (without having been explicitly training to do so).

Table 2: Question answering accuracy.

EMBEDDING	DIM	α	SCORE
NO HINTS	-	0.0	0.84%
BERT-REST	768	0.0	1.08%
SIF	768	0.7	3.14%
SIF	768	0.2	3.29%
USE	512	0.0	4.45%

Note that the results in Table 2 are not directly comparable with the reported accuracy of the 1.5 billion parameter GPT-2-1.5B (4.1%), since the "Yes/No" questions have been deliberately excluded in the experimental results above, since random chance would then add approximately 1.8% (of pure noise) to the results presented here. Adjusting the reported GPT-2 figures (downward) for this effect shows that the proposed model has higher performance for a much lower parameter count, even when using purely unsupervised training methods.

6 Discussion

As mentioned in Sutskever (2019), an online video in which Radford et al. (2017) is discussed, 'higher order' capabilities seem to appear in language-related models only if the size of the model is sufficient to have captured many the basic features of the underlying language, since knowing the basic words and structures is more important to a Language Modeling objective than higher order features like sentiment and story arc (for instance).

Being able to capture such higher order features provides a natural incentive to want to scale the training of language models to as large a number of parameters as possible. And undoubtedly there will be important and interesting results to come out of these efforts.

However, it is not at all clear that embedding factoids in neural network weights is a practical way of building intelligent systems. Even humans (built on a biological neural substrate) seem to reason about facts symbolically *despite* the processing being based in neurons.

The goal of this research is to explore how to interface the extremely effective aspects of models such as GPT-2 with more accessible sources of knowledge and planning.

By using the *human readable* output of a Language Model component to direct further information gathering (or, potentially, other activities), one might imagine the system would not only become more capable (without exponentially long training), but would also have an *internal dialogue* that would be human interpretable.

6.1 Further Work

Clearly, more experimentation is needed to understand how to improve the current system. Fortunately, that can be accomplished without a huge investment in hardware.

In terms of sentence embedding techniques, one additional method was investigated, so far without encouraging results : the generation of sentence embeddings from using an additional layer for the GPT-2-117M model it its initially untrained state. This deserves further work, given the findings of Wieting and Kiela (2019).

Also interesting is the potential for training a more specific retrieval/utilisation engine in a supervised manner, such as in Bapna and Firat (2019), and then expanding the domain across which retrieval is performed to encompass a much broader range of accessible facts without further training the model. However, this is slightly contrary to the goal herein of using purely unsupervised techniques.

Beyond these initial phases, though, there is the potential for the system to achieve some level of self-improvement. As was discussed in Radford et al. (2019), the GPT-2-1.5B model could not only answer some factoid questions, but it also had a good (self-) model of confidence in its answers[2]. This implies that if a trainable embedding component were included in *this* paper's architecture it might be trainable (in a fully self-supervised way) to improve its self-hinting, and thereby achieve a self-improving positive feedback loop.

Acknowledgments

The authors would like to thank Google for access to the TFRC TPU program which was used in training and fine-tuning models during experimentation for this paper.

References

Chris Alberti, Kenton Lee, and Michael Collins. 2019. A BERT baseline for the Natural Questions. *Computing Research Repository*, arXiv:1901.08634.

[2] "The probability GPT-2 assigns to its generated answers is well calibrated and GPT-2 has an accuracy of 63.1% on the 1% of questions it is most confident in."

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*.

Ankur Bapna and Orhan Firat. 2019. Non-parametric adaptation for neural machine translation. *Proceedings of the 2019 Conference of the North*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *Computing Research Repository*, arXiv:1803.11175.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *Computing Research Repository*, arXiv:1810.04805.

H Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics, Vol. 3, Speech Acts*. Academic Press, New York.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *Computing Research Repository*, arXiv:1904.09751.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. *Computing Research Repository*, arXiv:1704.01444.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Ilya Sutskever. 2019. Deep unsupervised learning.

John Wieting and Douwe Kiela. 2019. No training required: Exploring random encoders for sentence classification. *Computing Research Repository*, arXiv:1901.10444.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.