Tackling the Scientific Question Answering Using the Dataset: ARC

Manlin Zhang
Arizona State Uniersity
mzhan129@asu.edu

Chi Duan
Arizona State Uniersity
cduan1@asu.edu

Yilun Huang
Arizona State Uniersity
yhuan329@asu.edu

Fumin He
Arizona State Uniersity
fuminhe@asu.edu

Hong Guan
Arizona State Uniersity
hguan6@asu.edu

Abstract

There has been tremendous progress in natural language processing(NLP) technology in recent years, majority of which are using various neural network techniques. Along with the emerging NLP technologies, there also have been emerging various NLP data sets and tasks with higher level of difficulties as new challenges to demonstrate the performance limitations of these technologies. Those two aspects of NLP research, technology and data sets, are evolving quickly in a manner that better technology raises research communities' interest to create the more difficult data sets to test those models and more challenging data sets will drive researchers to development more innovative technologies. This process has expedited the NLP research achievements remarkably with the increasing capacity from computing power from GPUs. In this paper we focus on tackling the challenge posed by AI2 Reasoning Challenge (ARC) data set, which is one of the most challenging and newest question answering data set. This data set is consisted of 7,787 grade-school science questions and partitioned into Challenge Set and Easy Set. The tough part of tackling of this data set lies in searching for the relevant paragraph from the background knowledge. Current we plan to test reinforcement learning search based method and neural network based search method in our model. Our language model will be based on RoBERTa or BERT.

1 Introduction

In recent years NLP technology has made remarkable progress with the neural network based models after there has been breakthrough in the the learning of word representations or embedding (Mikolov et al., 2013a) and (Mikolov et al., 2013b). With the embedding as input, most of emerging NLP models are built with Recurrent neural networks(RNN) and Long-Short-Term-Memory(LSTM) neural networks. However the word embedding has its own limitation by the fact that it is not a context-aware process when words are embedded the into vectors. Then later the transduction based model Transformer (Vaswani et al., 2017) bring the NLP research into a new stage. More research are adopting the Transformer as building blocking in their models. The OpenAI GPT model (Radford et al., 2018) is built with sequences of left-to-right Transformers. One step further from OpenAI GPT, the BERT model (Devlin et al., 2019) is using bidirectional Transformer and can be further used as input to many other NLP tasks, enabling it to become a very widely adopted as a base building block in many research. In addition to BERT, the model RoBERTa (Liu et al., 2019) adds recipes to better train the BERT model. Evolving together with rapidly emerging new models, newer and more challenging data sets are also created to be stages for models to show their performances and limitations. Among the many dataset categories, the question answering and reasoning is a popular category, in which many data sets have been created, such as SQuAD1.0 (Rajpurkar et al., 2016) and SQuAD2.0 (Rajpurkar et al., 2018), SNLI (Bowman et al., 2015), CoQA (Reddy et al., 2018), HotpotQA (Yang et al., 2018), SearchQA (Dunn et al., 2017). In this paper we will focus on a particular question answering and reasoning data set: the AI2 Reasoning Challenge (ARC) (Clark et al., 2018) data set, which is consisted with grade school science questions and more difficulty than its predecessor question answering data sets. In phase 1, we are testing models: reinforcement learning search based model and neural network based search model. Currently we are still in the development phase. But we are expecting our model will achieve state-of-art performance.

2 Dataset and Task Description

2.1 Dataset Description

The ARC data set (Clark et al., 2018) contains 7,787 natural grade-school level multiple-choice SCIENCE questions. This data set's level of difficulty requires far more powerful knowledge and reasoning capability than ever before data sets such SQuAD1.0 (Rajpurkar et al., 2016) or SNLI (Bowman et al., 2015). The data set has a corpus as background knowledge. The corpus is using search engines with queries instantiated from hand-crafted templates and the corpus covers 99.8% of the questions' vocabularies (Clark et al., 2018). The data set also has two partitions: EASY Set and CHALLENGE Set. The Challenge Set contains only questions answer incorrectly by both a retrieval-based algorithm and a word co-occurence algorithm. Examples from both sets are as below:

1. *EASY: Which technology was developed most recently? (A) cellular telephone (correct) (B) television (C) refrigerator (D) airplane*

2. *CHALLENGE:What does photosynthesis produce that helps plants grow? (A) water (B) oxygen (C) protein (D) sugar (correct)*

And inside each set, it is also divided into train, test and development sets. The statistics regarding the ARC data set are summarized in Table 1. The corpus is used to to give background information in the training process. But the ARC challenge is not limited to this corpus as sole source of knowledge and at some level it could also be viewed as open book challenge meaning that it can use external knowledge from other sources than the corpus.

Subset	Questions	Total
Easy-Train Set	2251	
Easy-Test Set	2376	
Easy-Development Set	570	
Easy-Total		5197
Challenge-Train Set	1119	
Challenge-Test Set	1172	
Challenge-Development Set	299	
Easy-Total		2590

Table 1: The statistics of ARC data set.

3 Methods

3.1 Limitations of Previous Models

We observed some general problems in the current BERT-based NLP systems and in this course project, we are going to make our first step to tackle these problems. The first problem is that the BERT (Devlin et al., 2019) model has a limited length of processing a document (512 wordpieces). The second problem is that BERT models themselves have limited power of selecting and ranking. Let's consider an example, in the setting of OpenBookQA problem and multihop reasoning problems, a typical idea to incorporate external knowledge is to first extract external knowledge either from the given resources (e.g. the open book) or from a search engine (e.g. Lucene), and second, concatenate the question, options, and external knowledge in some way to form an input, and then fine-tune a BERT model using this input and a human labeled option. This method is likely to exceed the length limit of BERT and also doesn't fully consider the priority of the obtained external knowledge. We also found that to answer some of the questions, one can not consider the choices independently.

3.2 Proposed Methods

In this project, we used the AristoRoBERTav7 model as the baseline. We investigated multiple aspects of this kind of pretrained neural models, specifically, the architecture of the model, size of the model, external knowledge usage, and input/output formats. For this purpose, we designed a series of experiments. We summarize our experiments in Table 2.

Although we didn't have a systematic ablation study, we can present some important observations that can lead to some insights of the neural models and the ARC challenge.

3.3 Additional Background Knowledge

As mentioned in (Clark et al., 2018), the ARC corpus covers 99.8% of the question vocabularies and relates approximately 95% to the ARC Challenge questions. It is obvious that not all of the question can be answered based solely on the corpus . Thus external knowledge other than the corpus would help the model to improve accuracy. Even for questions than can be answered from with knowledge the corpus, the external knowledge still help to provide redundant information to that question ,which

Aspects	Controlled experiments
Input format	Compare separate choices and combined choices as inputs (using RoBERTa)
Model size	Compare various model size of the same architecture (T5 base vs. T5 3B)
Model architecture	Compare various model architectures of comparable model size (RoBERTa base vs. T5 base)
External knowledge	Compare with external knowledge and without external knowledge (using RoBERTa)

Table 2: Aspects of a model and controlled experiments

will increase the probability of the model accrediting to the right answer. Thus in our work, we also decide to use external knowledge. We plan to use the search engine *Lucene* and *Aristo* to obtain as much external knowledge as possible. The obtained external knowledge will be combine ARC corpus as background knowledge in our model.

4 Baseline Model Evaluation and Error Analysis

In phase 1, we only discuss the set up and error analysis of the base model. Our base model is named AristoRoBERTaV7 developed by Aristo team at Allen Institute for AI based on the RoBERTa-Large model (Liu et al., 2019). The base model achieved accuracy around 0.66 as claimed on the ARC data set submission leader board, which is a descent score. We also run the base model on our own machine and achieved similar score. In our run, we used batch size of 8, epoch of 4 and learning rate of 10^{-5}. Other setting are same as default as in RoBERTa model.

When we examine the errors in the testing set, we found that most errors are caused by the mismatching of the background knowledge with the questions and answers. The search algorithm in the model first match each answer choice with some background knowledge paragraph as fact and then the language model decides the probability of that choice. If an answer is matched with wrong fact, it is very likely the model will go wrong and give false choice. In our error analysis, we found that errors can be divided into the main categories listed in the following. The example of the sample wrong question could also be found in Appendix A.1.

1. Fail to extract useful information from the corpus to answer the question.
2. Noisy information extracted from the corpus.
3. Lack of knowledge concerning Time, physics, math and chemistry, but no useful information extracted from the corpus.
4. Lack of other background knowledge, common sense, but no useful information extracted from the corpus.
5. Fail to understand synonyms or antonyms.
6. Fail to do co-reference resolution.
7. Fail to answer questions that need to consider all options.
8. Fail to recognize the order of concepts.

Those errors emphasizes the fact that matching the correct and perceivable background fact in the model is the key causes to inaccuracies in the model.

5 Experiment

In all the experiments, we adopted the same principle that we used the train set and dev set to develop and fine-tune our model and finally run the model on the test set. We used four combined datasets as train and dev set, while we kept only the ARC challenge test set as our final test set. These four datasets are ARC train (1119 qs), ARC-Easy train (2251 qs), OpenBookQA train (4957 qs), Regents Living Environments train (665 qs).

5.1 The RoBERTa family

We combined the questions and the choices as one sentence and fed it to the model, and used a linear layer of size (hidden size, 5) on top of the [CLS] token, where the number five is the number of classes since we have five classes at most. To help the model to better distinguish different choices, we augmented the data by randomly shuffling the choices while keeping the correct choice as the answer.

We also use RoBERTa to investigate if external knowledge help improve the accuracy. In our experiments, we used OpenBookQA corpus as the knowledge base, because we found it has less noise compared to the ARC corpus. Borrowing ideas from REALM (Guu et al., 2020), we simplified the algorithm in three ways: the first is to use locality sensitive hashing (Gionis et al., 1999) to quickly store and query high dimension vectors; the second is to use the same RoBERTa model as both the classifier and embedding extractor for the [CLS] tokens; the third is to combine the question and knowledge by concatenating the vector output representation of [CLS] tokens instead of concatenating their raw text inputs as one sentence. These modifications from the original paper ease the need computing powers and memory, especially the GPU memory. However, we didn't find good results by applying all these tricks, which may due to our oversimplification.

Nevertheless, we show an example of knowledge pieces that we retrieved from the OpenBookQA corpus.

1. *question: ARCCH question: Mixing baking soda and vinegar makes the temperature of the solution decrease and release carbon dioxide. Which conclusion about this investigation is not valid? choice1: Mixing the chemicals caused them to absorb heat. choice2: A chemical reaction took place. choice3: New elements were formed. choice4: The procedure caused a gas to be formed.*

2. *Ten extracted knowledge pieces with its score after softmax*

    (a) knowledge sentence 0: A new moon happens once per revolution of the moon.
    score: 0.8764

    (b) knowledge sentence 1: Earth is our planet.
    score: 0.0486

    (c) knowledge sentence 2: Lake Erie is a body of water.
    score: 0.0144

    (d) knowledge sentence 3: The arctic is desolate.
    score: 0.0141

    (e) knowledge sentence 4: The American television miniseries about slavery is called Roots.
    score: 0.0095

    (f) knowledge sentence 5: Paleontologists study fossils like old animal feces.
    score: 0.0092

    (g) knowledge sentence 6: Squirrels eat edible flowers.
    score: 0.0082

    (h) knowledge sentence 7: Bats listen to their voice to guide them.
    score: 0.0078

    (i) knowledge sentence 8: Terns are cold weather animals.
    score: 0.0067

    (j) knowledge sentence 9: Some fish are carnivores and cannibals.
    score: 0.0050

In our RoBERTa experiments, we use the following hyperparameters; learning rate, 1e-5; batch size, 16, gradient accumulate steps, 2; weight decay, 0.01.

5.2 The T5 family

T5-base model and T5-3B model are used in our experiments. We treated the ARC task as a close-book exam without any external knowledge. We processed the inputs by concatenating the question and choices as the input text and the correct choice with its content as the target text. Here's an example of our instance, input text: ”ARCCH question: George wants to warm his hands quickly by rubbing them. Which skin surface will produce the most heat? choice1: dry palms choice2: wet palms choice3: palms covered with oil choice4: palms covered with lotion”; target text: ”choice1: dry palms”. Like the RoBERTa experiments, we augmented the training data by randomly shuffling the choices.

5.3 The BERT family

We also used the BERT model. The main model tested in our project is the BERT-Base model. The BERT model can take up to two text sentences separated by the [CLS] tokens inputs. The model's output layers can be adapted to various vector length (Devlin et al., 2019). In our experiment setup, we use each question as one text input and one of the answer choice as another text input. Thus we convert each instance in the original ARC data set into four instances.The instance of question with the correct answer text is marked as 1 and the instance of question with the incorrect answer text is marked as 0. The following gives an example of this conversion.

1. *Original Data: Which technology was developed most recently? (A) cellular telephone (correct) (B) television (C) refrigerator (D) airplane*

2. *Converted Data:*

   (a) text_a = Which technology was developed most recently?
   text_b = (A) cellular telephone
   label = 1

   (b) text_a = Which technology was developed most recently?
   text_b = (B) television
   label = 0

   (c) text_a = Which technology was developed most recently?
   text_b = (C) refrigerator
   label = 0

   (d) text_a = Which technology was developed most recently?
   text_b = (D) airplane
   label = 0

   After training, the model then predicts on the converted development and test data in the same manner. The prediction of 1 means the the answer is correct to the question and 0 means incorrect. However, predicting on the converted data set may give a situation when multiple answers are predicted correct or no answers are predicted correct. In this situation we choose answer with most probability to be correct for multiple situation and the answer with the least probability to be incorrect for the no answer situation. As a consequence of this operation, we will have one exact answer for each question. We also tested the data set using BERT-Medium model which has smaller model size than BERT-base and is less computationally expensive to train. But BERT-Base model achieved a better performance. We didn't get a chance to train a BERT-Large model due to limitations of computing resource available. However we can expect that the model accuracy can be further improved with a larger model.

6 Results

We use exact match to measure the accuracy of the performance and summarize our results statistics in Table 3.

Model	**Test accuracy**
RoBERTa-base without knowledge	42.1
RoBERTa-base with knowledge	25.6
T5-base	44.8
T5-3B	69.0
BERT-base	35.5

Table 3: ARC challenge test accuracy of various models

The unsatisfied results of our RoBERTa-base with knowledge model may due to our over-simplication of the knowledge retrieval process, particularly, one single RoBERTa model is burdened with both training representation of the knowledge sentence and classification. It may also due to the cold start problem of the retriever.

Here are examples of a correct prediction and an incorrect prediction for our T5-3B model:

1. *Correct Prediction:*

   arcch question: a star with twice the mass as the sun would? choice1: start its life cycle with a fission process. choice2: use its fuel source much more quickly. choice3: provide more solar flares. choice4: result in a nebula.

   Target: choice2: use its fuel source much more quickly.

   Prediction: choice2: use its fuel source much more quickly.

   Counted as Correct? True

2. *Incorrect Prediction:*

   arcch question: what advantage do some plants have over animals in a drought? choice1: plants can use underground water. choice2: plants can live without water. choice3: plants release oxygen. choice4: plants cannot move.

   Target: choice1: plants can use underground water.

   Prediction: choice2: plants can live without water.

   Counted as Correct? False

7 Conclusion

In this project, we empirically investigated multiple factors of the neural models that may influence the performance. Though it is not a systematic ablation study, we shown how different model size, neural net architecture, training strategies lead to performance difference. Specifically, we found that T5 has better performance over RoBERTa and BERT of comparable model size by a small margin. We also found that combining all choices with data augmentation is beneficial. More importantly, from the fact that the T5-3B model significantly outperform the T5-base model, we conclude our most important finding, a model with large size is able to encode knowledge.

8 Future Work

Although larger models are more likely to capture some knowledge in the world, it is still worthy to explore new methods that are more efficient. Our model with augmented knowledge seems to be one of the directions. Though the model didn't get high performance in our pilot experiments, we still consider it as a potential method to tackle the knowledge retrieval problem. In addition, we proposed an RL-based method in our phase 1 report. Unfortunately it was not implemented when we submit our project, yet it is a good potential solution to accurately identify relevant knowledge.

Acknowledgments

Each team member is actively involving in the project and we regard contributions from each member to be equal.

References

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.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Güney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *CoRR*, abs/1704.05179.

Aristides Gionis, Piotr Indyk, and Rajeev Motwani. 1999. Similarity search in high dimensions via hashing. In *VLDB*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *Technical report, OpenAI*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2018. Coqa: A conversational question answering challenge. *CoRR*, abs/1808.07042.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *CoRR*, abs/1809.09600.

A Appendices

A.1 Samples of wrong outputs from RoBERTa-base model

1. Sample Question 1:
   Question text: *"Which of the following can be found on both Earth and the moon?"*
   Correct Answer: *The correct answer is "A. Hills"*
   Error Reason:*We cannot find useful information from the paragraph.*
   label: *"A", "text": "Hills",*
   para/fact: *"As well as the Benedictine Abbey on the hill, he founded three other monasteries close to the town and he created on the hills of Down a city, both monastic and mercantile, of which both the mediaeval and the twentieth century citizens can be proud. As a result, Easter falls on the first day of the week (Sunday) after the first full moon following the spring equinox, and thus can be as early as March 22 and as late as April 25 [which would make it the second full moon after the equinox] (ibid., McGraw Hill, NY, 1967, pp. 1062-1063). It was midnight before we followed, and the moon, which we had watched rise over the hills, was so bright that a torch was unnecessary. The northern line follows the creek, except the Half Moon, a loop in front of Academia, which is included in Spruce Hill. Parking can be found along the hill and at the top on both the right and left. Shizuku on the hill 450x792 62K GIF drawn by; The painter which drawn this fan art said, it shows the following situations that Shizuku had pursued the cat Moon from the station , she lost sight at once, but she found it again on the hill. As well as the Benedictine Abbey on the hill, he founded three other monasteries close to the town and he created on the hills of Down a city, both monastic and mercantile, of which both the medieval and the twentieth century citizens can be proud. of the blessings which we derive from quot;the sun and moon, and the everlasting hills,quot; from the succession of the seasons and the produce of the earth; It can be found in the southern hills of both Kovisberg and Halenberg. Which of the following can be found in Sicily: hills, mountains, plains, lowlands?"*

2. Sample Question 2:
   Question text: *"A 0.20 kg softball travels 97 meters (m) south for 4.5 seconds (s). What piece of information distinguishes the velocity from the speed of the ball?"*
   Choices: *A. "The ball went south.", B. "The ball flew for 4.5 s.", C. "The ball traveled 97 m.", D. "The ball has a mass of 0.20 kg."*
   Correct Answer: *A*
   Error Answer:*B*
   Error Reason :*Lack of knowledge concerning Time, physics, math and chemistry, but no useful information extracted from the corpus.*

3. Sample Question 3:
   Question text: *"A student wishes to measure the rate of evaporation of a volatile liquid at room temperature. Which piece of equipment would be best to use in this investigation?"*
   Choices: *A. "balance", B. "microscope", C. "meter stick", D. "litmus paper"*
   Correct Answer: *A*
   Error Answer:*B*
   Error Reason :*Fail to understand synonyms or antonyms.*