Explore whether state-of-the-art pre-trained models in NLP can achieve general reading and reasoning abilities

Junyu Yin
School of Computer Science and Engineering, Nanyang Technological University
s210027@e.ntu.edu.sg

1 Motivation & Objectives

As far as we humans are concerned, in order to achieve good performance on reading comprehension tasks, reading through a given passage and understanding what it says are just the preliminary requirements. We can see that the texts in the reading comprehension tests from the elementary school level to the postgraduate examination level gradually changed from simple to complex, and the question types also changed from simple text extraction to complex logical reasoning. And in this process, we humans have developed some general reading comprehension abilities through continuous learning and training. As a result, it is natural to ask that whether a computer, or specifically an AI system, can also acquire this combination of natural language understanding and logical reasoning.

With recent advances in deep learning techniques, it seems possible to achieve human-level performance in certain language understanding tasks, and a surge of effort has been devoted to the machine comprehension task where people aim to construct a system with the ability to answer questions related to a document that it has to comprehend. Just like the human language test, the most common way to test whether an intelligent agent (a person or an AI system) can fully understand a piece of text is to require her/him/it to answer questions about the text. This is called Question Answering (QA).

For the simple QA in which the candidate answers are directly extracted from the context, many models can achieve excellent performance. And for complex QA which requires logic reasoning ability, many dataset have been developed in recent years and many researchers claimed that they have developed models with the logic reasoning ability.

However, I suspect that these AI models, while performing well on relevant complex datasets, do not actually understand the texts and do not acquire general logical reasoning capabilities at all. They just learn different implicit biases and features contained in different datasets, and use this information to obtain answers on test sets with the same data distribution. More specifically, suppose we train and test a model on a dataset based on graduate reading tests which require the ability to do complex reasoning and to understand more complex texts. If the model performs well and we assume that it has real reasoning ability, then the model should perform equally or better on lower-level reading tests. In other words, a person who can achieve good grades in reading comprehension at the graduate level is unlikely to perform worse in reading comprehension at the junior high school level.

In my opinion, even the current powerful pre-trained models are just learning the underlying biases of the dataset and cannot develop strong generalization abilities like we humans do. So in this project, I want to first (1) conduct experiments to see whether the current state-of-the-art pre-trained models which are trained and behave well on complex QA dataset have learned general reading and reasoning skills and if possible (2) to seek that whether I can improve the model performance by adding some new variant.

2 Methodology

2.1 NLP tasks

In this project, I plan to do the multiple choice question answering (MCQA) task, which refers to identifying a suitable answer from multiple candidates by estimating the matching score among the triple of the passage, question and answer, to check whether the models can learn general reasoning skills.

The input format of this task is a triple (C, Q, A) with C denoting the context, Q the question and

A the answers. Some concrete examples can be seen in 1, 2 and 3. And the output format is just a predicted answer.

2.2 Dataset Analysis

In this project, I currently plan to experiment with three datasets, RACE, LogiQA and ReClor. They are both MCQ type datasets with 4 options–only one of them is correct. Their details are as follows.

2.2.1 RACE

Collected from the English exams for middle and high school Chinese students in the age range between 12 to 18, RACE (Lai et al., 2017) consists of 27,933 passages and 97,687 questions generated by human experts (English instructors), and covers a variety of topics which are carefully designed for evaluating the students' ability in understanding and reasoning.

This dataset is further divided into two subgroups RACE-M, which denotes the middle school examinations, and RACE-H, which denotes high school examinations to distinguish the two subgroups with drastic difficulty gap.

To better understand this dataset, I list the subdivisions of questions under the reasoning category, including detail reasoning, whole-picture understanding, passage summarization, attitude analysis and world knowledge.

Since all the data is from the English test at the secondary level, I believe that only a moderate level of reading comprehension and a moderate level of logical reasoning are required to perform well on this dataset. To get a better sense of this, I show a sample from this dataset in 1.

2.2.2 LogiQA

LogiQA (Liu et al., 2020), which contains 8,678 paragraph-questions pairs, is constructed by collecting the logical comprehension problems from publically available questions of the National Civil Servants Examination of China, which are designed to test the civil servant candidates' critical thinking and problem solving. Its English version is translated from the Chinese original by professional English speakers. To ensure the quality of the translation, the authors of the original dataset also hired proofreaders to do further proofreading.

The dataset covers a wide range of logical reasoning types, including categorical reasoning, conditional reasoning, disjunctive reasoning, and conjunctive reasoning. Thus, a model needs to acquire strong reasoning ability to achieve good performance on this dataset. However, the passages of this dataset are not complicated at the grammatical level, or even relatively simple. Therefore, the requirements for reading comprehension of this dataset are not very high.

Therefore, I think we should achieve a moderate level of reading comprehension and a high level of logical reasoning to perform well on this dataset. To better understand this, I show a sample from this dataset in 2.

2.2.3 ReClor

ReClor (Yu et al., 2020) are constructed by collecting the reading comprehension problems in some high-level standardized tests, such as GMAT and LSAT. To be specific, it contains 6,138 data points, in which 91.22% are from actual exams of GMAT and LSAT while others are from high-quality practice exams.

The length of the context of ReClor is much shorter than RACE and is a litter longer than LogiQA and the length of answer options of ReClor is largest among these datasets. In addition to this, the texts in the ReClor dataset cover larger vocabularies (nearly 26,576) as well as more complex grammatical structures. This means that we need stronger reading comprehension skills to fully understand the text in this dataset.

In RACE, there are many redundant sentences in context to answer a question. However, in ReClor, every sentence in the context passages is important. And it covers much more types of reasoning skills whose percentages and descriptions are shown in 4. This means that high logical reasoning skills are required to correctly answer questions in ReClor.

What's more, the authors of the original dataset leveraged some techniques to find out the biased data points in ReClor and divided the whole dataset into two subsets named EASY and HARD respectively. The division method can refer to the original paper, and will not be repeated here.

So I do believe that we need a high level of reading comprehension ability and a high level of logical reasoning techniques to perform well on this dataset. To better understand this, I show a sample from this dataset in 3.

2.3 Neural Models & Method

In this project, my initial idea was to find a model that achieved good performance on all three datasets simultaneously. However, during the in-

Passage:
In a small village in England about 150 years ago, a mail coach was standing on the street. It didn't come to that village often. People had to pay a lot to get a letter. The person who sent the letter didn't have to pay the postage, while the receiver had to.
"Here's a letter for Miss Alice Brown," said the mailman.
" I'm Alice Brown," a girl of about 18 said in a low voice.
Alice looked at the envelope for a minute, and then handed it back to the mailman.
"I'm sorry I can't take it, I don't have enough money to pay it", she said.
A gentleman standing around were very sorry for her. Then he came up and paid the postage for her.
When the gentleman gave the letter to her, she said with a smile, " Thank you very much, This letter is from Tom. I'm going to marry him. He went to London to look for work. I've waited a long time for this letter, but now I don't need it, there is nothing in it."
"Really? How do you know that?" the gentleman said in surprise.
"He told me that he would put some signs on the envelope. Look, sir, this cross in the corner means that he is well and this circle means he has found work. That's good news."
The gentleman was Sir Rowland Hill. He didn't forgot Alice and her letter.
"The postage to be paid by the receiver has to be changed," he said to himself and had a good plan.
"The postage has to be much lower, what about a penny? And the person who sends the letter pays the postage. He has to buy a stamp and put it on the envelope." he said . The government accepted his plan. Then the first stamp was put out in 1840. It was called the "Penny Black". It had a picture of the Queen on it.

Questions:

1): The first postage stamp was made _.
A. in England B. in America C. by Alice D. in 1910

2): The girl handed the letter back to the mailman because _ .
A. she didn't know whose letter it was
B. she had no money to pay the postage
C. she received the letter but she didn't want to open it
D. she had already known what was written in the letter

3): We can know from Alice's words that _ .
A. Tom had told her what the signs meant before leaving
B. Alice was clever and could guess the meaning of the signs
C. Alice had put the signs on the envelope herself
D. Tom had put the signs as Alice had told him to

4): The idea of using stamps was thought of by _ .
A. the government
B. Sir Rowland Hill
C. Alice Brown
D. Tom

5): From the passage we know the high postage made _ .
A. people never send each other letters
B. lovers almost lose every touch with each other
C. people try their best to avoid paying it
D. receivers refuse to pay the coming letters

Answer: ADABC

Figure 1: An example in the RACE dataset.

P1: David, Jack and Mark are colleagues in a company. David supervises Jack, and Jack supervises Mark. David gets more salary than Jack.

Q: *What can be inferred from the above statements?*
A. Jack gets more salary than Mark.
B. David gets the same salary as Mark.
C. One employee supervises another who gets more salary than himself.
✓ **D. One employee supervises another who gets less salary than himself.**

P2: Our factory has multiple dormitory areas and workshops. None of the employees who live in dormitory area A are textile workers. We conclude that some employees working in workshop B do not live in dormitory area A.

Q: *What may be the missing premise of the above argument?*
A. Some textile workers do not work in workshop B.
B. Some employees working in workshop B are not textile workers.
✓ **C. Some textile workers work in workshop B.**
D. Some employees living in dormitory area A work in the workshop B.

Figure 2: An example in the LogiQA dataset.

Context:
In jurisdictions where use of headlights is optional when visibility is good, drivers who use headlights at all times are less likely to be involved in a collision than are drivers who use headlights only when visibility is poor. Yet Highway Safety Department records show that making use of headlights mandatory at all times does nothing to reduce the overall number of collisions.
Question: Which one of the following, if true, most helps to resolve the apparent discrepancy in the information above?
Options:
A. In jurisdictions where use of headlights is optional when visibility is good, one driver in four uses headlights for daytime driving in good weather.
B. Only very careful drivers use headlights when their use is not legally required.
C. The jurisdictions where use of headlights is mandatory at all times are those where daytime visibility is frequently poor.
D. A law making use of headlights mandatory at all times is not especially difficult to enforce.
Answer: B

Figure 3: An example in the ReClor dataset.

Type	Description
Necessary Assumptions (11.4%)	identify the claim that must be true or is required in order for the argument to work.
Sufficient Assumptions (3.0%)	identify a sufficient assumption, that is, an assumption that, if added to the argument, would make it logically valid.
Strengthen (9.4%)	identify information that would strengthen an argument
Weaken (11.3%)	identify information that would weaken an argument
Evaluation (1.3%)	identify information that would be useful to know to evaluate an argument
Implication (4.6%)	identify something that follows logically from a set of premises
Conclusion/Main Point (3.6%)	identify the conclusion/main point of a line of reasoning
Most Strongly Supported (5.6%)	find the choice that is most strongly supported by a stimulus
Explain or Resolve (8.4%)	identify information that would explain or resolve a situation
Principle (6.5%)	identify the principle, or find a situation that conforms to a principle, or match the principles
Dispute (3.0%)	identify or infer an issue in dispute
Technique (3.6%)	identify the technique used in the reasoning of an argument
Role (3.2%)	describe the individual role that a statement is playing in a larger argument
Identify a Flaw (11.7%)	identify a flaw in an argument's reasoning
Match Flaws (3.1%)	find a choice containing an argument that exhibits the same flaws as the passage's argument
Match the Structure (3.0%)	match the structure of an argument in a choice to the structure of the argument in the passage
Others (7.3%)	other types of questions which are not included by the above

Figure 4: The percentage and description of each logical reasoning type.

vestigation, I found that the state-of-the-art models on the three datasets adopted different architectures. The common part of these state-of-art models is using the large-scale pre-trained language models as their backbone. And to my best knowledge, I figured out that the RoBERTa (Liu et al., 2019) architecture was most used in published papers about this task. So I decide to use it as my neural model currently.

The current state-of-the-art results on machine reading are all based on the pre-trained models. Different from the traditional deep learning methods, pre-trained methods consider the context, the question and each candidate answer as one concatenated sentence, using a pre-trained contextualized embedding model to encode the sentence for calculating its score. Given four candidate answers, four concatenated sentences are constructed by pairing each candidate answer with the context and question, and the one with the highest model score is chosen as the answer. And for the implementation of pre-trained methods, we follow the HuggingFace implementation (Wolf et al., 2019).

To better understand what I am going to do in this exploration, I briefly discuss experimental design here. My purpose here is to verify that whether a model can achieve some general reading ability, rather than just learning proprietary latent features on different datasets. So I will use different combinations of the three datasets when fine-tuning the pretrained model, and test each combination strategy on three datasets at the same time. For example, if a model trained on the ReClor achieves good performance and we think that it acquire some general reading abilities, the model should also perform well on RACE, even if the model is not trained on RACE.

However, this is just my initial design idea, in the follow-up process, I will design more experiments to verify my conjecture. And if possible, I want to improve the pre-trained model to achieve better performance on all three datasets simultaneously.

3 Evaluation

3.1 Baselines

According to the above discussion, I use RoBERTa model which is trained and tested on the disjoint subsets of the same dataset as our baseline model here. The results of all three datasets are shown in 1.

3.2 Quantitative Analysis

For the MCQA task, we can just use accuracy as our evaluation metric. And further analysis only can be completed after the experiments are finished.

3.3 Qualitative Analysis

I would like to analyze the generalization ability of the model for reading comprehension and logical reasoning here. I'm more inclined to think that the

Dataset	Acc	Middle	High
RACE	83.2%	86.5%	81.3%
LogiQA	35.3%	-	-
ReClor	55.6%	75.5%	40.0%

Table 1: The first line comes from the RoBERTa paper, and the other two come from their dataset paper. RACE and ReClor are divided into 2 parts, we report their results respectively.

current model can't get this ability, but this analysis only can be done after I finish all the experiments.

References

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

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. *arXiv preprint arXiv:1907.11692*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. *arXiv preprint arXiv:2002.04326*.