MCQA by computational model in comparison to human behaviour

Stav Cohn
206025496
cohnstav@campus.technion.ac.il

Nurit Klimovitsky Maor
308453208
knurit@campus.technion.ac.il

Abstract

Multiple Choice Question Answering (MCQA) is widely used to study and measure reading comprehension in humans and in language models. The task of MCQA based on a given context text is a challenging task for language models. In traditional NLP research the main goal is to achieve models that select the correct answer with high accuracy scores. In this work we aim to compare the question answering abilities of a computational model to observed human behaviour. We utilize a RoBERTa model by fine-tuning it on RACE (Lai et al., 2017) and OneStopQA (Berzak et al., 2020). We run the fine-tuned model on OneStopQA and obtain the prediction distribution for each question. The data that is used as the observed human behaviour is the data that was gathered in (Berzak et al., 2020) using the crowd-sourcing platform Prolific (Pro). In this work we show the results analysis of the comparison between the model and the human responses.

1 Introduction

Reading comprehension can be measured by multiple choice question answering assessment. The task of Multiple Choice Question Answering (MCQA) includes 3 segments:

- A context text
- A question that corresponds to the context
- 4 options to choose an answer from - one correct answer and 3 distractors.

The chosen answer is usually expected to have correlation with reading comprehension level of the context text. A computational model that is trained to perform the MCQA task with sufficient similarity to the observed human behaviour can have paramount influence on many fields.

- Reducing the dependencies in data gathered by human surveys and studies. It can be done by using the model to get predictions that replace the need for human response.
- Identifying faulty questions that are used for assessment tests (such as SAT test or any other reading comprehension exams).
- Difficulty assessment of text and questions.
- Assessment of text simplification by comparing the predictions of a model in different levels of the context text.

2 Related Work

STRAC (Structured Annotations for Reading Comprehension) is an annotation framework for assessing reading comprehension with multiple choice questions, introduced in (Berzak et al., 2020). The STRAC framework is implemented in OneStopQA, a high-quality dataset for evaluation and analysis of reading comprehension in English. Our project is primarily based on this work. We used OneStopQA and data gathered from the experiments performed on Prolific.

RoBERTa The RoBERTa model was proposed in (Liu et al., 2019). It is based on Google's BERT model released in 2018. It builds on BERT and modifies key hyperparameters, removing the next-sentence pretraining objective and training with much larger mini-batches and learning rates. By utilizing RoBERTa we achieved higher accuracy score for predicting the correct label than by using the BERT architecture (Devlin et al., 2018)

Fine-tuned Roberta-large from HuggingFace (USP) is a fine-tuned model of Roberta-large applied on RACE dataset that was trained similarly to our model but required larger computational resources than the resources we had access to.

3 Data

The datasets used for this project are RACE and OneStopQA.

RACE The ReAding Comprehension Dataset from Examinations dataset (Lai et al., 2017) is a machine reading comprehension dataset consisting of 27,933 passages and 97,867 questions from English exams, targeting Chinese students aged 12–18. Each question is associated with 4 candidate answers, one of which is correct. Questions in RACE are specifically designed for testing human reading skills and are created by domain experts. RACE large dataset size was crucial for our training process, yet it also has some flaws. As shown in (Berzak et al., 2020) 47% of its questions can be guessed by machines without accessing the passage, and 18% are unanimously judged by humans as not having a unique correct answer.

OneStopQA provides an alternative test set for reading comprehension which alleviates these shortcomings. OneStopQA is a multiple choice reading comprehension dataset annotated according to the STARC (Structured Annotations for Reading Comprehension) scheme (Berzak et al., 2020). The reading materials are Guardian articles taken from the OneStopEnglish corpus. The dataset is consisted of 30 articles, 162 paragraphs 486 questions and a total of question-paragraph level pairs: 1,458. Every question has four possible answers. The first answer is the correct answer. Importantly, the correct answer typically does not appear verbatim in the passage. The second answer represents a misunderstanding of the critical information for answering the question correctly. The third answer refers to information in the passage that is not relevant for the question. The fourth distractor has no support in the passage. Due to this meticulous structure, this dataset is much smaller than RACE but the quality of the data is substantially higher.

Prolific.co is a crowd-sourcing platform for psychological research that provides access to a wide audience of participants from different cultural backgrounds on demand.

4 Model

4.1 Data Preprocessing

We decided to use RoBERTa-base with a head of Multiple Choice. The Multiple Choice head requires the input data to be in a specific format, hence a preprocessing pipeline is imperative. The preprocess pipeline follows some of the notions shown in (MC). We preprocessed the data in stages.

First, we adjusted the data from RACE and OneStopQA to a consistent template. The template holds all the information about the article, context, question, answer candidates and a label. Next, we tokenized the data. For each question item we created a set of 4 sequences in a format of context, separator token, question, another separator token and an answer candidate. Then we tokenized the sequence using the RoBERTa tokenizer. We truncated the tokenized sequences to have a length of 512 tokens by using a simple heuristic. The heuristic iteratively truncated the longer section (i.e. context/question/answer candidate) of a sequence until reaching a 512 sequence length. Lastly to get the data in the required format, we extracted the input ids and attention masks from the tokenizer for each of the 4 sequences of a question item.

4.2 Fine-tuning

The fine-tuning process was done by feeding the preprocessed question items to the model. For each question item, the model processes each of the 4 sequences separately. The model returns a single value for each of the 4 sequences and for the final prediction the 4 values are inserted into a softmax layer. The softmax layer returns the probability distribution of the sequences. This distribution is the probability distribution for the answer candidates assigned by the model.

We first fine-tuned the model on RACE. We ran the fine-tuning process on the RACE train dataset for 20 epochs, with batch size of 8, base learning rate of 1e-5 and the optimizer that was used was AdamW. We trained the model on 4 parallel GPUs. We evaluated the accuracy of the model on the RACE dev dataset. We saved the model configurations of the epoch that yielded the highest accuracy score on the RACE dev dataset - 74%.

To further improve the abilities of the model we then fine-tuned the model on OneStopQA in 4-fold split. We trained on 70% of the data and evaluated the accuracy on the other 30%. The predictions were then made on the 30% that were used for the evaluation. This process was repeated on each fold and the predictions were gathered to a total of 1719 questions. For all folds, the hyper-parameters used were the same as the hyper-parameters used

on RACE. The accuracy score for the folds was: 75.8%, 80.5%,79.6% and 76.4% (the size of the last fold was smaller than the rest). 10 epochs were run for each fold.

5 Analysis

5.1 Data preparation

The received model output contains predictions for 1719 context+question pairs. The prolificQA.tsv file contains questions that are not included in OneStopQA and vice versa. We cleaned, merged and aggregated the data to have one table. Each row contains the details of the question, the predictions of the model for each answer candidate and the distribution of the responses of the people that answered the question. Before aggregating the responses of the people on the questions the table contained 1728 rows. The merged table contained data for 577 questions.

We added additional columns to the data for further analysis.

- Distance columns - The distance between the probability of the human response and the predictions given by the model.
- Agreement columns - The probability of the human response to the answer that was selected by the model as the most probable answer.

5.2 Correct answers

Our goal was to analyse the similarities and differences in behaviour between humans and our model. To do so, we analysed the differences between the model predictions and the distribution of human response.

We first analyzed the ratio of the correct answers. As can be seen in Figure 1, for 79.93% of the questions, the model selected the correct answer, while people selected the correct answer for 87.72% of the questions.

We can analyze the same information from a different angle by taking the model's prediction of the correct answer and comparing it to the amount of correct answers given by human participants. In Figure 2 we see that when comparing the probability that is assigned by the model to the correct answer and the proportion of the people that answered correctly we get 75.72% for model prediction and 80.97% for human response - a smaller difference.

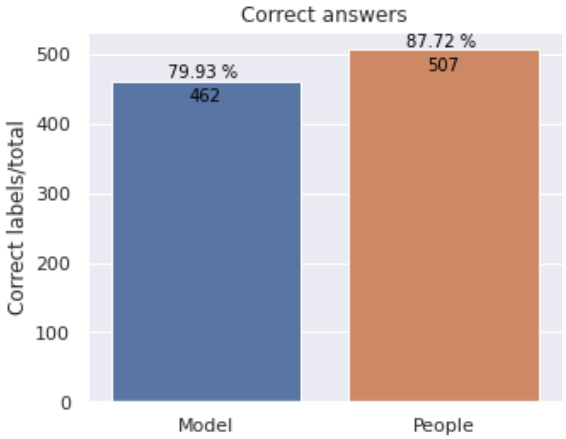

Figure 1: Correct answers from all questions

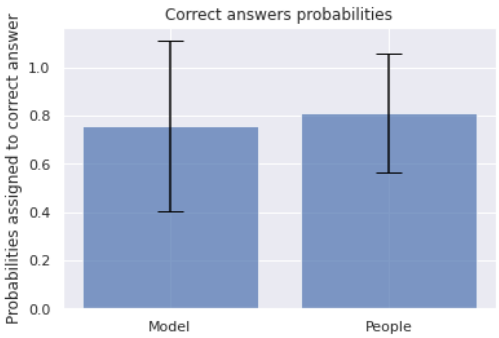

Figure 2: Probabilities of the correct answers

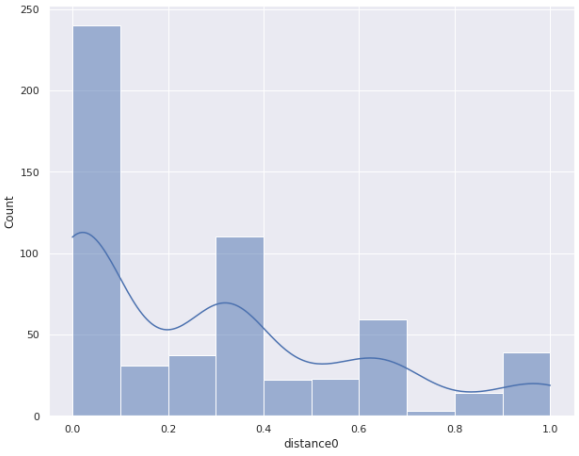

Figure 3: Histogram of the distance of the probabilities of the correct answers

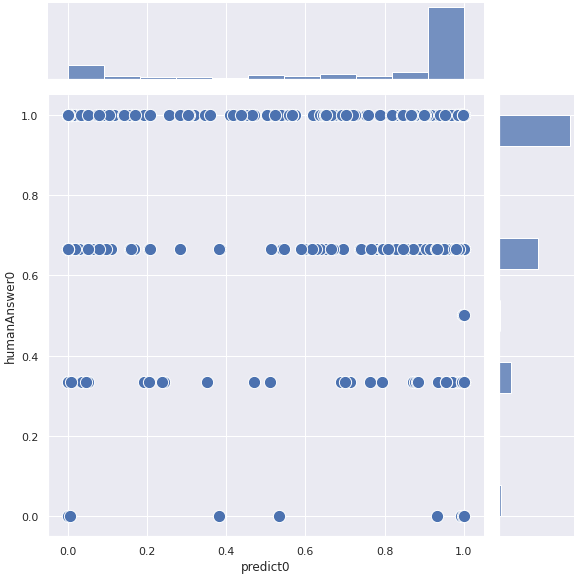

Figure 4: Each dot represents a question. The x value represents the probability the model assigned to the correct answer. The y value represents the proportion of the people that selected the correct answer for that question.

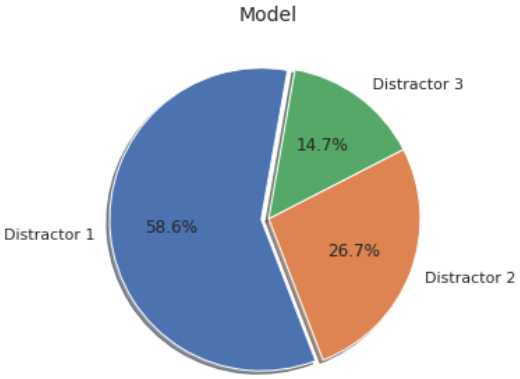

Figure 5: The proportion of questions in which each distractor is selected by the model from the questions in which any distractor was selected.

Figure 3 presents a histogram of the distance between the probability assigned to the correct answer by the model and the proportion of the people that selected the correct answer. We see that the largest bin belongs to the smallest distance - for 240 questions, which are 41.52% of the questions, the distance was smaller than 0.1.

In Figure 4 we see a scatter plot that shows the relationship between the human and model responses to the correct answers per question. X-axis is the probability that is assigned by the model to the correct answer while the Y-axis is the proportion of the people that selected the correct answer. The histogram in Figure 4 also shows that both the model and the people that responded to the questions answered most of the questions correctly and supports the claim following Figure 2.

5.3 Distribution of mistakes

There are some similarities between the distribution of the incorrect responses of the model (Figure 5) and the people (Figure 6). For both, when an incorrect answer is selected, for the most part, it is the first distractor - the distractor that belongs to the same span. The next most common mistake is the second distractor - the distractor that belongs to another span, for both the model and the human response. Last, with the smallest portion, is the third distractor - the distractor that does not have

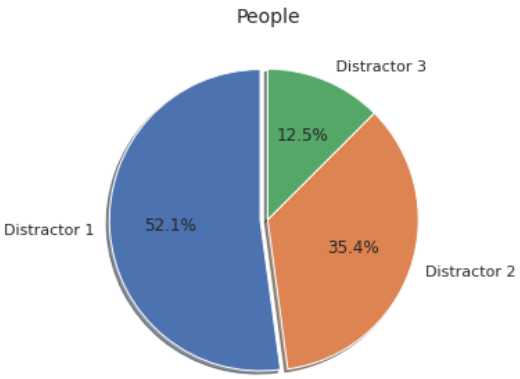

Figure 6: The proportion of human responses in which each distractor was selected.

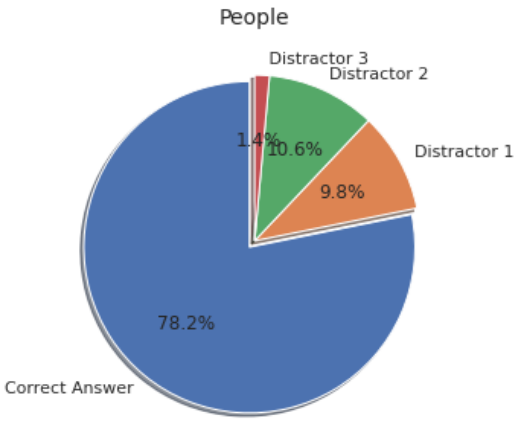

Figure 7: Proportions of the answers selected by people in the questions the model did not select the correct answer

a claim in the context, for both the model and the human response. This result corresponds with the STRAC scheme on which the analysed items were built.

Although the similarities are prominent there are some differences. There is a difference of 6.5% in the proportions of the first distractor between the model's selection and human response. There is a difference of 8.7% between the proportions of the second distractor. And lastly there is a difference of 2.2% between the proportions of the third distractor.

Figure 7 shows that 78% of the people responded correctly to questions in which the model selected a distractor.

5.4 Agreement

Another approach to the analysis of the differences between the model and humans would be to assess how similarly they respond to each question. To do so, we measure the level of agreement between the model and the people that have answered the question, we compare the proportions of the responses of the people to the answer that is selected by the model. The agreement was divided to 4 categories:

- No agreement - None of the people that responded to the question selected the same answer as the model.
- Minor agreement - at least $\frac{1}{3}$ of the people that responded to the question selected the same answer as the model.

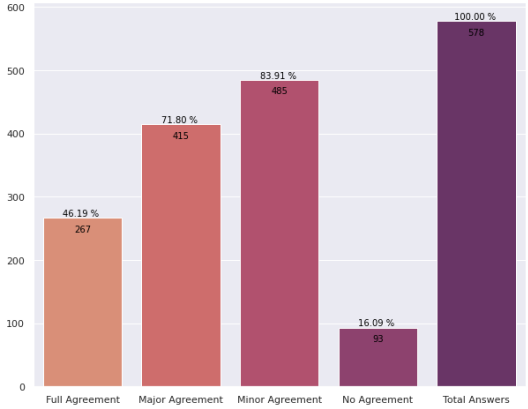

Figure 8: Agreement between the model and the people that responded to the questions

Agreement - Trained on RACE			
Level	**General**	**Correct**	**Incorrect**
Full	38.24%	58.01%	0%
Major	60.03%	88.71%	4.57%
Minor	72.49%	98.69%	21.83%
None	27.51%	1.31%	78.17%
Agreement - Trained on OneStopQA			
Level	**General**	**Correct**	**Incorrect**
Full	46.19%	57.58%	0%
Major	71.8%	88.53%	5.17%
Minor	83.91%	98.48%	25.86%
None	16.09%	1.52%	74.14%

Table 1: Agreement levels. General - all questions; Correct - questions in which the model selected the correct answer; Incorrect - questions in which the model selected a distractor. First subtable shows the agreement levels for the model that was trained on RACE only, second subtable shows the agreement levels for the model that was trained on RACE and then on OneStopQA

- Partial agreement - at least $\frac{2}{3}$ of the people that responded to the question selected the same answer as the model.
- Full agreement - all people that responded to the question selected the same answer as the model.

In Figure 8 we can see that for 71.8% of the questions most of the people agreed with the model on the selected answer (major agreement), for 46.19% of the questions all the people that responded to the question selected the same answer as the model (full agreement) and for 16.09% none of the people agreed with the model on the selected answer.

Table 1 shows the percentage of questions according to agreement categories.

As shown, for only 57.58% of the questions in which the model selected the correct answer there is an agreement between the model and **all** the respondents. This suggests that the model selected the correct answer in questions in which at least one person did not respond correctly, which means those questions were not trivial to people.

On the one hand, we can see from the Correct column that for the most part there is an agreement between the model and people's responses, when the model selects the correct answer. One the other, we can see from the Incorrect column that for the most part there is no agreement when the model does not select the correct answer. There is not even one question in which there is full agreement when the model selected a distractor.

5.5 Conclusions and final remarks

As shown, there are many similarities between the model and human responses, yet there are also some disagreements between the predictions and response distribution.

Further research can be done with more complex models (e.g. RoBERTa-large, LongFormer, etc.). A larger dataset with more information on human response on each question and on more questions could also improve the analysis and provide additional insights.

6 Experiments

6.1 Using BERT

A previous version of the model was based on the outdated BERT (before the merge to HuggingFace). This version yielded an accuracy score of 58% on RACE alone. This result was not significant enough to move forward with this model to the analysis stage. Consequently, the code was changed - the model used for the predictions was replaced with the latest version of RoBERTa and the needed adaptations to support the input format required by RoBERTa were added.

6.2 Fine-tuning on RACE only

A few key numbers for comparing to the analysis previously shown:

- Ratio of correct labeling - 65.92%
- Correct answer probabilities average - 63.39%
- Distractors selection proportions in questions in which the model selected a distractor
  - Distractor 1 - 44.2%
  - Distractor 2 - 33.5%
  - Distractor 3 - 22.3%
- proportion of human responses to questions in which the model selected a distractor
  - Correct answer - 79%
  - Distractor 1 - 10%
  - Distractor 2 - 9.1%
  - Distractor 3 - 1.9%

Level/Answer	0	1	2	3
Elementary	453	80	25	15
Intermediate	450	80	24	19
Advanced	444	86	26	17

Table 2: Rows - text difficulty as tagged in OneStopQA; Columns - the number of predictions per label

For most values, when comparing models abilities to human responses, the values of the model that was fine-tuned both on RACE and OneStopQA, were closer to the measured human responses. Further more, from table 1 we conclude that the general agreement has improved significantly after fine-tuning on OneStopQA. This implies that RACE is not enough to fine-tune a model that can respond sufficiently to higher quality data such as the data in OneStopQA.

6.3 Metric for quality of simplified text

As stated in the introduction, one possible use case for a model that responds to multiple choice questions similarly to humans, is evaluation of text simplifications. Although the model did not exhibit the similarity that is required to safely assume it can be used as a tool for such task, we were still interested to see if there is some correlation between the responses of the model to the level of the text. We did not have a meaningful conclusion out of Table 2, but we find it interesting and some similar analysis can be helpful for models that will be developed in the future.

7 Our experience

Our first attempt was to implement a RoBERTA model based on the platform Google Colab Pro using tutorials. As it turned out there were not many resources/tutorials available for this task. We were surprised to find out that the state of the art approach is to use models that work on repetitions of

the context and questions (by the number of answer candidates). The resources that were required to accomplish this task were larger than we initially expected. As a result the NLP lab's resources were assigned to us and we got to experience the advantages and disadvantages of working with remote servers that belong to our faculty.

7.1 Updating legacy project

After a few different attempts to create a pipeline from scratch, we received the legacy project that utilized the a legacy BERT model (before the merge of PyTorch's Pretrained and HuggingFace). Exploring the project was enlightening and we've enjoyed updating and expanding it to support the needs of our research.

7.2 Analysis

Although the analysis was interesting to perform, we think the data from Prolific is lacking as it is based on 3 responses for each questions. We believe the results might be completely different with data that is based on more responses to each question.

Acknowledgments

Throughout this entire project we have received incredible support and guidance from Dr. Yevgeni Berzak, Omer Shubi and Refael Tikochinski it was a pleasure working with you and learning from you.

References

Bertformultiplechoice and swag dataset example. https://github.com/rodgzilla/pytorch-pretrained-BERT/blob/multiple-choice-code/examples/run_swag.py. Accessed: 2022-05-20.

Prolific.co. https://www.prolific.co/.

Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. 2020. STARC: Structured annotations for reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5726–5735, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

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*.

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.

LIAMF USP. A fine-tuned model of roberta-large applied on race. https://huggingface.co/LIAMF-USP/roberta-large-finetuned-race. Accessed: 2022-05-10.