Reading Comprehension based Question Answering Model

IT412 - Natural Language Processing

Anand Pol	Abhishek Shah	Sana Baid	Shivangi Gajjar	Preet Amin
202011006	*202011017*	*202011019*	*202011023*	*201801051*
M.Tech ICT - ML	*M.Tech ICT - ML*	*M.Tech ICT - ML*	*M.Tech ICT - ML*	*B.Tech ICT*
DA-IICT, Gandhinagar	DA-IICT, Gandhinagar	DA-IICT, Gandhinagar	DA-IICT, Gandhinagar	DA-IICT, Gandhinagar

***Abstract*—Language Modelling, Machine Translation, Information Retrieval, Text Categorization, Text Summarization, et cetera, have been active research areas in Natural Language Processing. Reading Comprehension-based question answering, which usually comes under Machine Reading Comprehension tasks, has been in the spotlight for quite a long time. In this project, we attempted the RC-based Q&A Language Modelling technique, tried some state-of-art methods, achieved comparable results, and even surpassed baseline results for some models. Relevant details about the ReClor dataset, its structure, models used are described in subsequent sections. Also, the contributions of each team member are described in detail at the end.**

***Index Terms*—Reading Comprehension, ReClor, RoBERTa**

I. Introduction

Every year numerous amount of entrance examination takes place for graduate admissions. With the increase in number of candidates, subjective kind of examinations are less preferred due to evaluation complexities. Hence, easy to evaluate yet effective genres of questions are asked in these examinations. Reading Comprehension (RC) based examination is one such genre, in which a portion of text is provided and few questions related to the text are asked to the candidate. These questions can be either subjective type or objective. However, objective type questions are more preferred by well known entrance examinations like LSAT, CAT, and many more. The generation of manual answer key for RC based examinations is a time consuming task and is prone to human error.

In order to expedite the mentioned process, we aim to build a system that generates answer key for these examinations. The basic idea of the model is not only to make the machine read the textual content, but also understand the context behind it. Once the context is clear to the machine, it need to be able to choose correct answer to each Multiple Choice Question (MCQ) related to the text. It is a many-to-one classification model, with input as a textual sequence and output as probabilities of options for each MCQ. Many state-of-the-art architectures like transformer [5], BERT [3], RoBerta [4] and GPT [2] have been proposed for this problem with outstanding results. RoBerta-large and RoBerta-base language models are used for the experiments for this project.

The rest of document is organised as follows: *Section II* discusses about the ReClor dataset and its format. *Section III* contains the description of architecture used for the model, training process, testing process and K-fold validation. *Section IV* shows the obtained results and current leaderboard rankings in competition [1]. Individual member contributions are included in *Section V*. *Section VI* contains conclusion of the whole project.

II. Dataset Description

ReClor [6] dataset is used for this project. The dataset is created by collecting the questions from the various competitive exams like GRE, CAT, GMET and so on. The dataset contains 5138 total samples. And 1000 samples are available for the test. The data set is collected from the various types of the questions such as Necessary Assumptions, Sufficient Assumptions, Strengthen, Weaken, Evaluation, Implication, Conclusion/Main Point, Most Strongly Supported, Explain or Resolve, Principle, Dispute, Role, Match Flaws, Identify a Flaw, Technique, Match the Structure and others.

A. Format of the data

The data is available in the JSON format. The fields of a individual sample contains Context, question, answers, label and corresponding sample id. Sample data is shown in the Figure 1.

III. Methodology

A. Data Processing

The input data is available in JSON format. The extracted data is a list of dictionaries with the following keys: Context, Example Id, Ending, Questions and labels. Using custom function these lists are converted to objects. Data cleaning

Type: Conclusion/Main Point
Definition: identify the conclusion/main point of a line of reasoning
Context:
Whether or not one can rightfully call a person' s faithfulness a virtue depends in part on the object of that person' s faithfulness. Virtues are by definition praiseworthy, which is why no one considers resentment virtuous, even though it is in fact a kind of faithfulness – faithfulness to hatreds or animosities.
Question: Which one of the following most accurately expresses the overall conclusion drawn in the argument?
Options:
A. The object of a person's faithfulness partially determines whether or not that faithfulness is virtuous.
B. Virtuous behavior is praiseworthy by definition.
C. Resentment should not be considered a virtuous emotion.
D. Behavior that emerges from hatred or animosity cannot be called virtuous.
Answer: A

Type: Evaluation
Definition: identify information that would be useful to know to evaluate an argument
Context:
George: Some scientists say that global warming will occur because people are releasing large amounts of carbon dioxide into the atmosphere by burning trees and fossil fuels. We can see, though, that the predicted warming is occurring already. In the middle of last winter, we had a month of springlike weather in our area, and this fall, because of unusually mild temperatures, the leaves on our town' s trees were three weeks late in turning color.
Question: Which one of the following would it be most relevant to investigate in evaluating the conclusion of George's argument?
Options:
A. whether air pollution is causing some trees in the area to lose their leaves
B. what proportion of global emissions of carbon dioxide is due to the burning of trees by humans
C. whether unusually warm weather is occurring elsewhere on the globe more frequently than before
D. when leaves on the trees in the town usually change color
Answer: C

Fig. 1. Few Sample of various types of questions included in ReClor Dataset

operations, namely, removal of punctuation and digits are performed on these objects after which tokens are generated using the Roberta tokenizer. Lastly, features masks are generated from these tokens and fed to the models for the training process.

B. Training Process

Context, Questions and respective four options are given as an input to the model. The model gives individual probabilities for the available four options as a output.Final answer can be predicted as the highest among the probabilities for the individual options. Inputs are passed through tokenizer. The output of tokenizer are fed to the desired Transformer Model (RobertaBase , RobertaLarge , etc.). Generated intermediate outputs are passed to the fully connected dense layer. At the end, softmax layer is applied for getting the individual probabilities for the available four option. Refer Fig. 2 for more details.

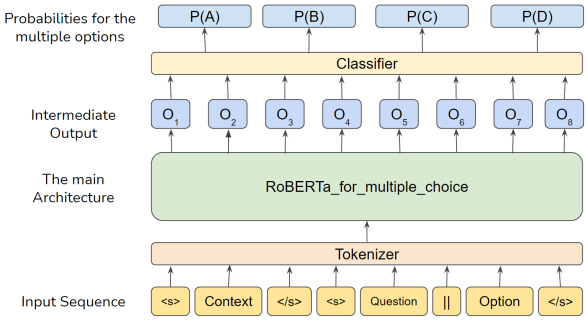

Fig. 2. Architecture used in the project

C. Testing Process

Accuracy is used as a metric for evaluation. Accuracy is obtained by comparing the predicted option with the corresponding ground truth value. Accuracy on test dataset is obtained by submitting the predictions on Test Data in a live competition [1]. After submission, we obtain accuracy for Test, Test-E as well as Test-H category, where Test is entire test dataset. Test-E stands for Easy and Test-H stands for Hard complexities based questions. Test-E and Test-H dataset is created by competition holder themselves. Refer Table IV for the test accuracy.

D. K-Fold Validation

Training and testing on one single train-validation set gave results as shown in Table II. The results can still be improved, so we tried K-Fold Validation. Initially, we combined both train and validation dataset, and created five different combinations of combined dataset. Then a Roberta-large model is trained and tested over five such Train-Val pairs and its results are as shown in Table III. Results show that after performing K-Fold validation, outcomes are improved and we got a test accuracy better than before.

IV. Results

All the results are generated on the GeForce-RTX 2080 Ti GPU and Xenon-based CPU with the batch-size of four. The test results are obtained by submitting the predictions on the ongoing competition [1]. Test-E and Test-H are defined in the competition itself. Current rank on the leader board based on the major categories are as follows.

- Based on Test: 13^{th}
- Based on Test-E: 7^{th}
- Based on Test-H: 11^{th}

In Table I, The baseline results of the Roberta-base and Roberta-large are shown. These results are taken from [6]. All the obtained results after fine-tuning the Roberta-base and Roberta-large included in the Table II. Results of the K-Fold validation can be found from Table III. All the codes for the implementation can be found from this github repository. https://github.com/anandpol98/ReClor-Reading-Comprehension-based-Question-Answering

ReClor	Roberta-base	Roberta-large
Val	55.0%	62.6%
Test	48.5%	55.6%
Test-E	71.1%	75.5%
Test-H	30.7%	40.0%

TABLE I
Baseline Results

Fold	Val	Test	Test-E	Test-H
Fold 1	**90.08 %**	54.30 %	75.00 %	38.04 %
Fold 2	55.06 %	53.00 %	70.00 %	39.64 %
Fold 3	57.59 %	55.20 %	73.40 %	40.89 %
Fold 4	56.42 %	55.30 %	**75.23 %**	39.64 %
Fold 5	56.23 %	**56.90 %**	74.54 %	**43.04 %**

TABLE III
Results obtained by performing K-Fold Validation using Roberta-large Architecture

ReClor	Roberta-base	Roberta-large	% Increment
Train	91.05%	97.33%	**6.89%**
Val	49.90%	54.90%	**10.01%**
Test	52.60%	63.6%	**20.91%**
Test-E	72.5%	78.41%	**8.15%**
Test-H	32.14%	36.43%	**13.34%**

TABLE II
RESULTS AFTER TUNING ROBERTA-BASE AND ROBERTA-LARGE. INCREMENT COLUMN SHOWS THE PERCENTAGE INCREMENT FROM ROBERTA-BASE TO ROBERTA-LARGE

V. CONTRIBUTIONS

We have divided the whole project into three parts (i.e. Literature survey and problem identification, Implementation, Report Writing). The whole project was done by all the members with equal contribution. We have divided our responsibilities for each of the project component. For the literature survey and problem identification, we have divided the selected papers among us and then collaboratively finalise the topic for the project.

Component	Responcible Persons
Literature Survey and Finding Problem Statement	All members
Data pre-processing and Extract features	Preet, Shivangi
Custom Code for $Roberta_{base}$	Sana, Preet
Custom Code for $Roberta_{large}$	Shivangi, Sana
Performing K-Fold Validation $Roberta_{large}$	Abhishek, Anand
Training and Testing	Anand, Abhishek
Report Writing	All members

TABLE IV
INDIVIDUAL CONTRIBUTIONS

For the implementation purpose, we have divided the whole task into the sub tasks. The involvement of the person to the particular sub-task(s) are shown in the Table II. For the report writing, all members have contributed equally.

VI. CONCLUSION

In this paper, Roberta architectures are successfully tuned and implemented for Reading Comprehension based QA task. With the fine tuning of Roberta-large model, we were able to surpass the baseline results. To improve it further, K-Fold validation was performed to increase the model performance. The results show that it helps to improve the performance of the model significantly. With the help of the K-fold validation and fine-tuning, we are able to secure 13^{th} position globally on the ongoing live challenge and achieved a notable 2.33% gain from the baseline results.

REFERENCES

[1] Evaluating state of the art in ai. evalai. (n.d.). retrieved october 28, 2021, from https://eval.ai/web/challenges/challenge-page/503/overview.

[2] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

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