Data enrichment of Binary Classification as an Effective Approach to Multi-Choice Question Answering

Pablo Torasso
Politecnico of Turin
Torino, Italy
S303196@studenti.polito.it

***Abstract*—We illustrate data enrichment with experimental results on refactoring of multichoice question answering tasks (MCQA) as a series of binary classifications, method proposed in the paper *Two is Better than Many? Binary Classification as an Effective Approach to Multi-Choice Question Answering* [1]. We introduce the score proposed in the original paper, the TEAM score (Two is bEtter thAn Many) , and we aim to show through experimental results that classifying (question, true answer) as positive instances and (question, false answer) as negative instances is significantly more effective than scoring each (question, answer) pair normalized over all the pairs, and then selecting the answer from the pair that yield the highest score. We show the efficacy of the proposed approach in question answering and non-English prediction of masked entity. The source code of the proposed approach is available at : https://github.com/Elbarbons/Data-enrichment-of-the-TEAM-method.**

***Index Terms*—TEAM, multi-choice question answering, QA, MCQA, transformers**

I. Introduction

The paper we consider was published in EMNLP (Empirical Methods in Natural Language Processing) 2022 conference by D. Ghosal, N. Majumder, R. Mihalcea, S. Poria; in this publication was firstly theorized and experimentally proved the efficiency of new binary classification model for multi-choice question answering (MCQA).

The results presented in the original paper are promising and suggest that the TEAM method, under certain conditions, can improve the accuracy of a MCQA pipeline. We focused our efforts to further enrich with experimental results the thesis proposed in the mentioned above document, by considering two previously untested datasets composed by two different task, reading comprehension dataset and masked entity prediction.

II. Methodology

Let q be a question for which multiple answer choices $A = \{a_1, ..., a_n\}$ are given. Optionally, some context c can be given to add key information to correctly answer the question. The purpose of the task is to identify the correct answer a_k from the answer set A. For the considered CSQA.hi dataset used in this paper, the question q is not provided, and can be implicitly assumed as: *What is the masked word in the given context ?.*

A. Score vs TEAM

In both methods each input sequence is constructed by concatenation of the question q, context c, and one possible answer choice a. The sequences are independently encoded through a pre-trained transformer language model such as RoBERTa or DeBERTa.

While the Score-based method (Score) generates a single s_i prediction for each input sequence and use his normalized distribution to evaluate the cross-entropy loss, where the loss is equivalent to the cross-entropy loss in a n-class classification setup, the classification-based method TEAM generates two unnormalized negative and positive scores for each input sequence and use their normalized distribution to obtain probabilities of negative and positive classes in order to calculate a specific loss function [1].

During inference, the answer with the highest positive class probability is choose as the predicted answer. We will confirm later in the section IV that the TEAM method outperform the Score method across both studied datasets for the same choiche of transformer models.

III. Experimental Dataset

We considered the following datasets:
RACE dataset(Lai et al. in, 2021). The ReAding Comprehension dataset from Examinations (RACE) dataset is a machine reading comprehension dataset consisting of passages and questions from English exams, targeting Chinese students aged 12-18. RACE consists of two subsets, RACE-M and RACE-H, from middle school and high school exams, respectively. RACE-M has 28,293 questions and RACE-H has 69,574. Each question is associated with 4 candidate answers, one of which is correct. [2]
CSQA.hi dataset(Kakwani et al. in, 2020). [3] The CSQA.hi dataset is part of the IndicGLUE natural language understanding benchmark for Indian languages. It contains a wide variety of tasks and covers 11 major Indian languages - as, bn, gu, hi, kn, ml, mr, or, pa, ta, te. It is a collection of hindi context answer pairs in which the task is to correctly guess the masked word in the context by selecting the correct answer between the given ones. We chose to benchmark this dataset with the TEAM method to analyze how the method performs with

Dataset	Istance
Race	**Question**: In which country is the prom called a ”formal”? **Choiche 1:** America . . . **Choiche 4:**Australia.
CSQA	**Question(Translated)**: Satimal MASK is a village in Vansda taluka, one of the five talukas . . . . **Choice 1:** Panchayatghar $\cdots$ **Choice 4:**$India$

Table I
ILLUSTRATION OF THE DATASETS USED IN THIS WORK.

non-English data and to confirm that the method is language-agnostic, as in the original paper only English dataset were considered.

IV. RESULTS

We use the DeBERTa Base and XLM-RoBERTa Base models to further benchmark the two previously introduced Score and TEAM methods across the experimental datasets. The Table 2II reports the accuracy score for the test set and compares accuracy scores between the two methodologies.

It is important to notice that the reported accuracy scores are not obtained with the top performing models such as DeBERTA-V3-XLarge due to hardware limits of the operating machine used to run the project, but the focus point behind this paper and the analysis of these metrics is the differences of accuracy performance between the two methods with same initial and input conditions, such as input dataset and language model.

Our finding is that the TEAM method improves over the Score method for both the datasets.

RACE: In this reading comprehension dataset from examinations the TEAM method performs better than the Score method for the reading comprehension question answering task. The first method outperformed the second method by around *3%* score on the test set with the DeBERTa language model.
Indic Glue: The proposed method is also better in the hindi masked entity prediction task with the usage of the XLM-RoBERTa language model. It achieved a score of *79%*, 3.4% better to respect of the Score method.

Model	Method	RACE	CSQA
DeBERTa-V3-Base	SCORE	71.42	
	TEAM	**74.00**	
XLM-RoBERTa-Base	SCORE	75.67	
	TEAM		**79.00**

Table II
ACCURACY ON THE TEST SPLIT OF THE DATASETS

V. ANALYSIS OF THE RESULTS

As reported in the original paper [1], the lexical similarity between the proposed answers in each instance of the dataset plays a key role in determining which method performs better; the TEAM method’s developers state that for the low to medium similarity datasets the TEAM algorithm performs the best, instead when a high lexical similarity is observed in the input data the Score algorithm performs better.

As shown in the table III, we reported the lexical similarity for the two datasets evaluated with BLEU [4] and ROUGE-L [5]. Both datasets has low lexical similarity between the right answer of a given question and wrong answers. This very low lexical similarity is translated in a better performance for the TEAM algorithm respect to the Score algorithm.

Dataset	BLEU	ROUGE-L	Δ
RACE	1.5	21.4	2.58
CSQA	~ 0	~ 0	

Table III
AVERAGE SIMILARITY BETWEEN CORRECT AND INCORRECT ANSWER CHOICES IN TEST SET. NUMBERS ARE SHOWN ON A SCALE OF 0-100. Δ INDICATE DIFFERENCE IN ACCURACY BETWEEN TEAM AND SCORE

The developers suppose that the softmax activation captures better the difference between the very similar correct and incorrect choices over the answers in the Score method. This aspect is not captured in the TEAM method, as sequences corresponding to the correct and incorrect choices are separately classified as positive or negative. Thus, is supposed that the Score method performs better when the answer choices are very similar. [1]

Also, as discussed in the section IV, the TEAM method performed very well with the non-English language dataset. As in the *Two is Better than Many?* paper the experimentation is conducted only on English data, we wanted to prove that the conditions of better results for the TEAM method are language agnostic. Indeed, the TEAM algorithm starts on top of the output given by the multilingual language model, so the results also confirm that the input language does not influence the TEAM accuracy performance.

VI. CONCLUSIONS

In this paper, we extended the *Two is Better than Many? Binary Classification as an Effective Approach to Multi-Choice Question Answering* with more experimental results and further confirm that the proposed binary classification method is a valid alternative to address multi-choice question answering task under certain circumstances.

VII. LIMITATIONS

Although the TEAM method produced encouraging results in the MCQA task, it lacks the flexibility of open-ended question answering and needs to assume the availability of a candidate answers.

H. Experimental Details

We trained the score-based and classification-based model with the AdamW optimizer with a learning rate of 1e-6, 3e-6, 5e-6. The models were trained for 8 epochs. The DeBERTa-V3-Base and XLM-RoBERTa-Base models were used to satisfy and not exceed the calculation power of the hardware used (reported in)

I. Computational Resources

We used a single NVIDIA RTX 3060 Ti GPU for ours experiments. Training takes between 30 minutes and 4 hours for the different datasets depending on what percentage of the training data is given as input to the model.

J. Dataset Details

The used datasets are available through the huggingface datasets hub. The number of MCQA instances in the training, validation and test set of the various datasets are shown in Table 3.

Dataset	Train	Validation	Test
RACE	25421	1436	1436
CSQA	18286	2387	2387

Table IV
Number of MCQA instances in the train, validation, and test set for the experimental datasets.

K. Splitting of the CSQA.hi dataset

As the CSQA.hi data retrieved from the Indic Glue dataset is not splitted in the training, evaluation and test datasets we manually splitted it with following percentages: 80% traning set, 10% evaluation set and 10% test set.

References

[1] D. Ghosal, N. Majumder, R. Mihalcea, S. Poria, Two is Better than Many? Binary Classification as an Effective Approach to Multi-Choice Question Answering, EMNLP 2022.

[2] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, Eduard Hovy, RACE: Large-scale ReAding Comprehension Dataset From Examinations, EMNLP 2017.

[3] Divyanshu Kakwani,Anoop Kunchukuttan, Satish Golla, Gokul N.C.,Avik Bhattacharyya ,Mitesh M. Khapra , Pratyush Kumar, IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages, Findings of the Association for Computational Linguistics 2020.

[4] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, Bleu: a Method for Automatic Evaluation of Machine Translation, 2002.

[5] DChin-Yew Lin, ROUGE: A Package for Automatic Evaluation of Summaries, 2004.