Learning to Classify the Wrong Answers for Multiple Choice Question Answering  
(Student Abstract)

**Hyeondey Kim, Pascale Fung**  
Center for Artificial Intelligence Research (CAiRE)  
The Hong Kong University of Science and Technology  
hdkimaa@connect.ust.hk

**Abstract**

Multiple-Choice Question Answering (MCQA) is the most challenging area of Machine Reading Comprehension (MRC) and Question Answering (QA), since it not only requires natural language understanding, but also problem-solving techniques. We propose a novel method, Wrong Answer Ensemble (WAE), which can be applied to various MCQA tasks easily. To improve performance of MCQA tasks, humans intuitively exclude unlikely options to solve the MCQA problem. Mimicking this strategy, we train our model with the wrong answer loss and correct answer loss to generalize the features of our model, and exclude likely but wrong options. An experiment on a dialogue-based examination dataset shows the effectiveness of our approach. Our method improves the results on a fine-tuned transformer by 2.7%.

**Introduction**

Machine Reading Comprehension (MRC) and Question Answering (QA) can be applied to many tasks, such as customer service and information search. It is one of the most obvious yardsticks for the evaluation of machine intelligence. Nowadays, on surface-level information extraction-based QA tasks, such as SQuAD (Rajpurkar et al. 2016), the machine already outperforms humans.

However, on the examination based QA task, the machine still has a long way to go to achieve human-level performance. One of the challenges of Multiple Choice Question Answering (MCQA), in particular, is that the machine needs to classify the one and only correct answer among the possible options. Even if the machine knows that one option makes great sense, there is no guarantee that it will be the correct answer. To avoid this problem, the machine must learn how to take the exam.

Mainly focusing on the DREAM task (Sun et al. 2019), we propose a new method, a wrong answer ensemble with a fine-tuned transformer. For humans, one of the common strategies for multiple-choice questions is the elimination method, i.e., getting rid of all of the wrong options. This method not only reduces the possibility of choosing the wrong option but also makes it easier to figure out the correct answer. Imitating this strategy, we consider both right options and wrong options together by combining the correct answer finding model and wrong answer finding model, in order to avoid choosing completely incorrect answers in the end. The main contributions of our work are as follows.

- We introduce a simple but effective method which is model-agnostic, and thus can be easily adapted to various MCQA tasks.
- The proposed model improves the performance of a fine-tuned transformer (Devlin et al. 2018) by 2.7%.

**Methodology**

Commonsense, multiple sentence level understanding and logical thinking are the key to success in any natural language understanding task. To achieve such abilities, we leverage the pre-trained language model BERT (Devlin et al. 2018). Because BERT is trained with next sentence prediction, it shows great performance on many natural language understanding tasks, especially on MRC.

**Fine-tuned Transformer**

In order to fine-tune the transformer, we adopt one of the BERT models implemented for the RACE dataset (Pan et al. 2019). We link the given questions $q$, option $o$, and corresponding dialogue $d$ into the [CLS] and [SEP] BERT token. In addition, we input sequences [CLS]$d$q[SEP]$o$[SEP] and separate the context, question and option with segmentation embedding.

**Wrong Answer Ensemble model**

Inspired by the elimination method, we propose a novel Wrong Answer Ensemble (WAE) model for MCQA. We implement the same fine-tuned transformer for the different tasks, and leverage a cross-entropy loss function for the Correct Answer (CA) model and a binary-cross-entropy loss function for the Wrong Answer (WA) model, respectively.

For the CA model, let $\hat{y}$ be the predictions, and we leverage the cross-entropy loss function. Hence, for each question, let $y$ be the labels. Then the loss value $\text{Loss}_{\text{correct}}$ for
In order to train our WA model to classify wrong answers, we label wrong options as ‘1’ and the right options as ‘0’. For the WA model, we deploy a sigmoid function for each wrong answer logits to make the prediction value of incorrect options in the model (i.e., the correct answer) take negative values. Let $x$ be the logit value of the WA model. Then

$$
\text{Sigmoid}(x) = \frac{e^x}{e^x + 1}.
$$

Then, let $\hat{y} = \text{Sigmoid}(x)$, and $y$ is the vector representation of the labels. The loss value for the WA model is

$$
\text{Loss}_{\text{wrong}} = - \sum y \cdot \log \hat{y} + (1 - y) \cdot \log 1 - \hat{y}.
$$

To merge the CA and WA models, let the final prediction of the CA model be $P_c$ and the final prediction of the WA model be $P_w$. Both prediction ranges are from -10 to 10. Instead of simply subtracting the logit values of the WA model from the logit value of the CA model, we leverage a simple linear regression to find the best weight value: $w$. Therefore, the final prediction is

$$
\text{Prediction} = \arg\max(p_c - w \cdot p_w)
$$

We achieve the best result when $w$ is 5.2 for the BERT-large model, and 2.2 for the BERT-base model. When the WA model’s performance approaches the CA model’s performance, the $w$ value increases.

### Experiment

We train our models for 24 epochs, with a batch size of 8, and a 1.5e-5 learning rate with the BERT pre-trained model.

### Evaluation

Our baseline model is the BERT-based fine-tuned transformer. Table 1 shows the evaluation results on the DREAM test set. We can see that WAE improves the BERT-base model by 1.7%, and the BERT-large model by 2.7%.

### Error Analysis

Our WA model shows improved performance on all of the question types with the BERT-large model. However, with the BERT-base model, on the matching type question, WAE shows lower performance than the CA model. One of the possible explanations is that matching is the easiest type of question in the DREAM dataset, normally contains one obvious option for the correct answer. Hence, WAE fails to improve the performance with the BERT-base model. Moreover, the two tasks, CA and WA, are very similar, but different due to the sigmoid function which is added to the WA model. The WA task requires classifying two wrong answers together and giving positive values to both of the wrong options. Therefore, it shows better performance on detecting wrong options when those options are similar to the right option. By training the model with two different labels, WAE generalizes the features of the model, and avoids choosing likely but wrong options.

### References


