Holding BERTs Hand Through Multiple Choice

Martin Bakač, Martin Josip Kocijan, Nikola Kraljević

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{Martin.Bakac,Martin-Josip.Kocijan,Nikola.Kraljevic2}@fer.hr

Abstract

Question answering tasks have long been a central challenge within NLP. This paper presents our approach to tackling the Sentence Puzzle task in the SemEval-2024 Task 9 competition: "BRAIN-TEASER: A Novel Task Defying Common Sense." Diverging from the conventional multi-class classification framework, we introduce a novel pairwise comparison methodology. Instead of forcing the model to choose the correct answer from four options, we reformulate the task into multiple decisions, evaluating triplets of answers at a time. This approach aims to make the model more robust to distractory and enhance the model's reasoning capabilities and accuracy. We compare results using this method and previously explored reformulations of multiple choice as a multi-class classification problem and a series of binary classification choices. We aim to find out which of these ways of *asking* BERT a multiple choice question works best on what task – SemEval SentencePuzzle and WordPuzzle. Our results showcase that the pairwise approach yields worse results than the multi-class classification approach, but better than the binary classification approach on both problems.

1. Introduction

The competition, named "*SemEval 2024 BRAINTEASER: A Novel Task Defying Common Sense*" described in the paper (Jiang et al., 2023) is here to push modern models to their limits, putting them to a quite adversarial setting. The paper makes a distinction between vertical and lateral thinking. Vertical thinking, also known as linear, logical, or convergent, is mostly a sequential and analytical process, requiring direct memory recall or a few logical steps to come to a sensible conclusion. Lateral thinking, also known as "thinking outside the box," is more of a divergent and creative process, where the question might not make sense when reading it at first, and to come up with an answer, you need to explore multiple angles. An even more adversarial task for transformer-based models is answering questions that require lateral thinking. State-of-the-art models such as ChatGPT show an accuracy of 60%, humans show a 90% accuracy, while random guessing gets close to 25% accuracy as there are four answer candidates for each question (Jiang et al., 2023). In this paper, we tackle the "BRAIN-TEASER" puzzles with an alternative approach which we believe will make it easier for transformer models to reason about all the possible answers, focusing on BERT-like models (Devlin et al., 2019). The puzzles are structured as multiple choice questions with one question being *"None of the above"*. Our goal is to compare different ways of formulating a multiple choice question by fine tuning BERT-like models on these formulations. In Section 3. we will describe the form of the two types of puzzles. Our approach, the reasoning behind it and other formulations of the multiple choice problem will be described in Section 4.. Results of our experiment will be presented in Section 5. and Section 6. respectively.

2. Related Work

Reasoning in NLP is a hot topic, with transformer-based models in the center of attention. It is no secret that transformer-based models are far from perfect, and it takes only one session of asking questions about a topic one is well versed in to see inconsistencies, but also be amazed at times. Knowing where a model fails is important and leads to improvement. Many benchmarks have been made with the purpose of challenging models with more intricate reasoning, like "Commonsense QA" (Talmor et al., 2019), a dataset consisting of questions that are easy for humans and require no prior knowledge, like a specific document or context, just common sense. Similarly, the "BRAIN-TEASER" (Jiang et al., 2023) is a benchmark dataset consisting of questions which are constructed in such a way that it is required to consider multiple approaches when answering them. As this was a competition dataset, there were some previous works that tackled this problem. A team of researchers from the National Technical University of Athens published their submission (Panagiotopoulos et al., 2023) named "*Transformer Models for Lateral Thinking Puzzles*". They showed promising results, significantly outperforming baselines of 60% reported in (Jiang et al., 2023). Their approach consisted of fine tuning models with a straightforward approach, treating the problem as a multi-class classification task. Additionally, they performed a transformation of the problem to a binary classification problem. The transformation took form of taking each question with four candidate answers and transforming it to three questions which required a binary label, signaling if the answer was correct or not [1]. Their results showed that the transformation was not quite useful, as the same models had somewhat worse accuracies when the transformation was applied. This inspired us to explore another transformation of the problem, as we believe that subjecting a model to only one candidate answer takes away information. When one is answering a question with candidate answers, it is a good idea to eliminate candidates that are somewhat obviously wrong. That is why we believe an approach where we force a model to choose between two candidate answers at a time might be helpful, this is described in detail in Section 4.

[1] The fourth candidate answer was always "None of the above" so they discarded it.

Table 1: A **word puzzle** question(sub-task B). The correct answer is in **bold**.

Question	Candidates
What part of London is in France?	**The letter N.** The letter O. The letter L. None of the above.

Table 2: A **sentence puzzle** examples(sub-task A). The correct answer is in **bold**.

Question	Candidates
A man shaves everyday, yet keeps his beard long.	**He is a barber.** He wants to maintain his appearance. He wants his girlfriend to buy him a razor. None of the above.

3. Dataset

As we've mentioned before the "BRAINTEASER" dataset consists of two types of puzzles, "Sentence puzzles" and "Word puzzles". Word puzzles focus more on letter composition, defying the original meaning of the words at hand, shown in Table 1. The second type of puzzles we find more interesting, so called "Sentence puzzles", shown in Table 2. The three step process of constructing the "BRAINTEASER" dataset is described in great detail in (Jiang et al., 2023). The process consists of data collection (web scraping), distractor sampling and generating reconstruction examples. As some pre-trained models could have these questions in their training corpus, steps were made to ensure that the questions were novel for the models (Su et al., 2019). The data set consists of original data, semantic reconstruction data, context reconstruction data. The semantic reconstruction used an open-source rephrasing[2] tool and human annotators as quality control to rephrase the questions. The context reconstruction process used GPT-4 to initially shift the context of the question, as well as human annotators. In the following sections we will refer to these distinct types of questions as **Ori.**, **Sem.** and **Con.** for original questions and semantic or context reconstruction respectivley. Lastly, the data was split into $507/120/120$ questions for the *train/test/validation* datasets respectivley.

4. Methods

In this section we will discuss the different approaches to multiple choice question asking and describes the training pipeline. The source code is available on github[3].

[2] https://quillbot.com/

[3] https://github.com/MBakac/BRAINTEASER

4.1. Reframing multiple choice

This section will discuss the main point of this paper – different formulations of a multiple choice question for BERT. We first describe the already explored multiclass classification and binary classification approaches and introduce the pairwise classification approach.

4.1.1. Multiclass classification

For this approach, the question and answers are concatenated together and the problem is framed as multiclass classification where the classes are answers A, B, C and D (None of the above). The main advantage of this approach is that the model sees all the answers in every pass.

4.1.2. Binary classification

In reframing the problem as a binary classification task we take every question and answer pair and label it 1 if the answer is the correct one, 0 otherwise. When determening the final choice answer for each question, we group the pairs by question and if the model outputs 1 for only one answer we chose that answer, otherwise D (None of the above) is chosen. One advantage of this is that we artificially have more traning examples, i.e. question answer pairs.

4.1.3. Pairwise approach

For this, new approach, we preform a transformation of the questions in the following way. Each tuple of a question and its candidate answers

$$(Q, A, B, C, D)$$

gets transformed into three tuples:

$$(Q, A, B, D),$$
$$(Q, A, C, D),$$
$$(Q, B, C, D).$$

In case the correct answer was C, the ground truth label in the tuple (Q, A, B, D) is D (None of the above). For each set of touples, *votes* for each answer are counted and the most voted answer is chosen. We aim to see if this approach that samples distractors multiple times for each example yields better results.

4.2. Training process

In line with the approach taken by (Panagiotopoulos et al., 2023) we load a pretrained BERT-like model from huggingface. The data and evaluation function is modified according to the reformulation and a BERT-like model is fine-tuned to the data.

For this work we decided to use general pretrained models RoBERTa-large[4] (Liu et al., 2019) and DeBERTa-v3-base[5] (He et al., 2021) as our goal was to compare different formulations, not to achieve the highest score. For future work, this could be tried with other, less general, fine-tuned BERT-like models.

[4] https://huggingface.co/FacebookAI/roberta-large

[5] https://huggingface.co/microsoft/deberta-v3-base

Table 3: Sentence puzzle

System	Original	Semantic	Context	Ori. + Sem.	Ori. + Sem. + Con.	Overall
Multiclass classification problem						
Human	.907	.907	.944	.907	.889	.920
ChatGPT	.608	.593	.679	.507	.397	.627
RoBERTa-L	.435	.402	.464	.330	.201	.434
microsoft/deberta-v3-base	.775	.775	.700	.775	.675	.750
FacebookAI/roberta-large	.850	.850	.775	.850	.700	.825
Binary classification problem						
microsoft/deberta-v3-base	.650	.650	.525	.625	.625	.608
FacebookAI/roberta-large	.125	.125	.125	.125	.125	.125
Pairwise						
microsoft/deberta-v3-base	.650	.570	.650	.650	.525	.623
Facebook/roberta-large	.675	.700	.725	.650	.600	.700

Table 4: Word puzzle

System	Original	Semantic	Context	Ori. + Sem.	Ori. + Sem. + Con.	Overall
Multiclass classification problem						
Human	.917	.917	.917	.917	.900	.917
ChatGPT	.561	.524	.518	.439	.292	.535
RoBERTa-L	.195	.195	.232	.146	.061	.207
FacebookAI/deberta-v3-base	.625	.688	.625	.594	.375	.646
FacebookAI/roberta-large	.312	.375	.438	.250	.188	.375
Binary classification problem						
microsoft/deberta-v3-base	.000	.000	.000	.000	.000	.000
FacebookAI/roberta-large	.000	.000	.000	.000	.000	.000
Pairwise						
microsoft/deberta-v3-base	.562	.469	.500	.469	.250	.510
FacebookAI/roberta-large	.250	.281	.406	.219	.125	.250

5. Results

We compared the three described formulations of multiple choice answering across two BERT variantas and look at the model accuracy overall and specifically for *semantic* and *context* subsets separated for the Sentence puzzle and Word puzzle subtasks (Table 3 and Table 4). Instance- and group-based performance metrics for our models as well as the human, ChatGPT and RoBERTa-L baselines are presented. The results are presented in terms of accuracy. The *Original*, *Semantic* and *Context* columns in the tables represent the results on respective reconstructions in the dataset while *Ori. + Sem* and *Ori. + Sem. + Con.* represent accuracy scores across multiple reconstructions of the same question.

Looking at Table 3, it is evident that the metrics show that the accuracy is homogenous with regards to the set of questions. In order words, if an original question is answered correctly, it is likely that the semantically and contextually reconstructed questions will be answered correctly as well. We take this as a good sign, because it signifies how the models are able to reason about the context behind the questions.

Table 4 shows that our models exhibit unhomogenous accuracy, or that answering an original question accurately does not make it substantially likelier to answer the semantically and contextually reconstructed questions correctly. This is because Word puzzles are harder to solve by locating the context behind the answer, as evidenced by ChatGPT's baseline performing worse in the Sentence puzzle subtask.

The calculation for group metrics in the pairwise approach was performed in a characteristic way. This was done in a matter-of-factly way by implementing a simple voting mechanism. Since the results are of acceptable quality, the voting system is justified. However, implementing other voting systems, such as a weighted voting system, which is outside of the scope of this paper, could improve the results in the future.

Context and semnatic reconstruction results

Zero accuracy in binary classification formulation A clear outliar in Table 4 are the accuracy results in in the binarcy classification formulation of multiple choice. This is due to the *harsh* criteria for deciding on an answer with this approach where out of three binary question only one has to be a *A*, *B* or *C* and the other two *D (None of the above)*. This, combined with the rather small dataset and, also the word puzzle task being overall harder (our models yielded overall worse accuracy scores for this task) meant that there were no cases where this approach worked, hence accuracy zero. Furthermore, we note that the models have been trained to have a strong bias towards answering negatively to all instances. This is partially caused by the fact that the dataset used to train the models for binary classification is imbalanced, with a ratio of 1:3 between positive and negative instances. This is a significant issue, as the models are not able to learn the patterns of the positive instances, which are the ones that we are interested in finding out. The same poor results for this formulation were obtained by Panagiotopoulos et. al. (Panagiotopoulos et al., 2023).

As a whole, the results of the study are remarkable in that human accuracy is within reach of our best-performing models by less than 10%. Despite the popularity of ChatGPT, it is not the best model for this task. Our models have managed to the baseline reference paper ChatGPT model by well over 10% in overall accuracy and by 30% in group accuracy (Jiang et al., 2023). This is a significant improvement over the current state-of-the-art models, which is a testament to the power of our approach.

6. Conclusion

To conclude, we have implemented and tested a new method for reformulating multiple choice questions. Our pairwise approach performed better than the binary classification reformulation but worse than the multiclass classification reformulation. For further research we suggest looking into other possible reformulations, expanding the dataset and using more fine-tuned models.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Yifan Jiang, Filip Ilievski, Kaixin Ma, and Zhivar Sourati. 2023. Brainteaser: Lateral thinking puzzles for large language models.

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.

Ioannis Panagiotopoulos, Giorgos Filandrianos, Maria Lymperaiou, and Giorgos Stamou. 2023. Ails-ntua at semeval-2024 task 9: Cracking brain teasers: Transformer models for lateral thinking puzzles.

Dan Su, Yan Xu, Genta Indra Winata, Peng Xu, Hyeondey Kim, Zihan Liu, and Pascale Fung. 2019. Generalizing question answering system with pre-trained language model fine-tuning. In Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen, editors, *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 203–211, Hong Kong, China, November. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge.