ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning

Abstract

Recent powerful pre-trained language models have achieved remarkable performance on most of the popular datasets for reading comprehension. It is time to introduce more challenging datasets to push the development of this field towards more comprehensive reasoning of text. In this paper, we introduce a new Reading Comprehension dataset requiring logical reasoning (ReClor) extracted from standardized graduate admission examinations. As earlier studies suggest, human-annotated datasets usually contain biases, which are often exploited by models to achieve high accuracy without truly understanding the text. In order to comprehensively evaluate the logical reasoning ability of models on ReClor, we propose to identify biased data points and separate them into EASY set while the rest as HARD set. Empirical results show that the state-of-the-art models have an outstanding ability to capture biases contained in the dataset with high accuracy on EASY set. However, they struggle on HARD set with poor performance near that of random guess, indicating more research is needed to essentially enhance the logical reasoning ability of current models.

Introduction

Machine reading comprehension (MRC) is a fundamental task in Natural Language Processing, which requires models to understand a body of text and answer a particular question related to the context. With success of unsupervised representation learning in NLP, language pre-training based models such as GPT-2 BIBREF0, BERT BIBREF1, XLNet BIBREF2 and RoBERTa BIBREF3 have achieved nearly saturated performance on most of the popular MRC datasets BIBREF4, BIBREF5, BIBREF6, BIBREF7. It is time to challenge state-of-the-art models with more difficult reading comprehension tasks and move a step forward to more comprehensive analysis and reasoning over text BIBREF8.

In natural language understanding, logical reasoning is an important ability to examine, analyze and critically evaluate arguments as they occur in ordinary language according to the definition from Law School Admission BIBREF9. It is a significant component of human intelligence and is essential in negotiation, debate and writing etc. However, existing reading comprehension datasets have none or merely a small amount of data requiring logical reasoning, e.g., 0% in MCTest dataset BIBREF10 and 1.2% in SQuAD BIBREF4 according to BIBREF11. One related task is natural language inference, which requires models to label the logical relationships of sentence pairs. However, this task only considers three types of simple logical relationships and only needs reasoning at sentence-level. To push the development of models in logical reasoning from simple logical relationship classification to multiple complicated logical reasoning and from sentence-level to passage-level, it is necessary to introduce a reading comprehension dataset targeting logical reasoning.

A typical example of logical reasoning questions is shown in Table TABREF5. Similar to the format of multiple-choice reading comprehension datasets BIBREF10, BIBREF5, it contains a context, a question and four options with only one right answer. To answer the question in this example, readers need to identify the logical connections between the lines to pinpoint the conflict, then understand each of the options and select an option that solves the conflict. Human minds need extensive training and practice to get used to complex reasoning, and it will take immense efforts for crowdsourcing workers to design such logical reasoning questions. Inspired by the datasets extracted from standardized examinations BIBREF5, BIBREF12, we build a dataset by selecting such logical reasoning questions from standardized exams such as GMAT and LSAT . We finally collect 6,138 pieces of logical reasoning questions, which constitute a Reading Comprehension dataset requiring logical reasoning (ReClor).

Human-annotated datasets usually contain biases BIBREF13, BIBREF14, BIBREF15, BIBREF16, BIBREF17, BIBREF18, which are often exploited by neural network models as shortcut solutions to achieve high testing accuracy. For data points whose options can be selected correctly without knowing the contexts and questions, we classify them as biased ones. In order to fully assess the logical reasoning ability of the models, we propose to identify the biased data points and group them as EASY set, and put the rest into HARD set. Based on our experiments on these separate sets, we find that even the state-of-the-art models can only perform well on EASY set and struggle on HARD set as shown in Figure FIGREF4. This phenomenon shows that current models can well capture the biases in the dataset but lack the ability to understand the text and reason based on connections between the lines. On the other hand, human beings perform similarly on both the EASY and HARD set. It is thus observed that there is still a long way to go to equip models with true logical reasoning ability.

The contributions of our paper are two-fold. First, we introduce ReClor, a new reading comprehension dataset requiring logical reasoning. We use option-only-input baselines trained with different random seeds to identify the data points with biases in the testing set, and group them as EASY set, with the rest as HARD set to facilitate comprehensive evaluation. Second, we evaluate several state-of-the-art models on ReClor and find these pre-trained language models can perform well on EASY set but struggle on the HARD set. This indicates although current models are good at exploiting biases in the dataset, they are far from capable of performing real logical reasoning yet.

Related Work

Reading Comprehension Datasets. A variety of reading comprehension datasets have been introduced to promote the development of this field. MCTest BIBREF10 is a dataset with 2,000 multiple-choice reading comprehension questions about fictional stories in the format similar to ReClor. BIBREF4 proposed

SQuAD dataset, which contains 107,785 question-answer pairs on 536 Wikipedia articles. The authors manually labeled 192 examples of the dataset and found that the examples mainly require reasoning of lexical or syntactic variation. In an analysis of the above-mentioned datasets, BIBREF11 found that none of questions requiring logical reasoning in MCTest dataset BIBREF10 and only 1.2% in SQuAD dataset BIBREF4. BIBREF5 introduced RACE dataset by collecting the English exams for middle and high school Chinese students in the age range between 12 to 18. They hired crowd workers on Amazon Mechanical Turk to label the reasoning type of 500 samples in the dataset and show that around 70 % of the samples are in the category of word matching, paraphrasing or single-sentence reasoning. To encourage progress on deeper comprehension of language, more reading comprehension datasets requiring more complicated reasoning types are introduced, such as iterative reasoning about the narrative of a story BIBREF20, multi-hop reasoning across multiple sentences BIBREF21 and multiple documents BIBREF22, commonsense knowledge reasoning BIBREF23, BIBREF24, BIBREF25 and numerical discrete reasoning over paragraphs BIBREF8. However, to the best of our knowledge, although there are some datasets targeting logical reasoning in other NLP tasks mentioned in the next section, there is no dataset targeting evaluating logical reasoning in reading comprehension task. This work introduces a new dataset to fill this gap.

Logical Reasoning in NLP. There are several tasks and datasets introduced to investigate logical reasoning in NLP. The task of natural language inference, also known as recognizing textual entailment BIBREF26, BIBREF27, BIBREF28, BIBREF29, BIBREF30 requires models to take a pair of sentence as input and classify their relationship types, i.e., Entailment, Neutral, or Contradiction. SNLI BIBREF31 and MultiNLI BIBREF32 datasets are proposed for this task. However, this task only focuses on sentence-level logical relationship reasoning and the relationships are limited to only a few types. Another task related to logical reasoning in NLP is argument reasoning comprehension task introduced by BIBREF33 with a dataset of this task. Given an argument with a claim and a premise, this task aims to select the correct implicit warrant from two options. Although the task is on passage-level logical reasoning, it is limited to only one logical reasoning type, i.e., identifying warrants. ReClor and the proposed task integrate various logical reasoning types into reading comprehension, with the aim to promote the development of models in logical reasoning not only from sentence-level to passage-level, but also from simple logical reasoning types to the complicated diverse ones.

Datasets from Examinations. There have been several datasets extracted from human standardized examinations in NLP, such as RACE dataset BIBREF5 mentioned above. Besides, NTCIR QA Lab BIBREF34 offers comparative evaluation for solving real-world university entrance exam questions; The dataset of CLEF QA Entrance Exams Task BIBREF35 is extracted from standardized English examinations for university admission in Japan; ARC dataset BIBREF12 consists of 7,787 science questions targeting student grade level, ranging from 3rd grade to 9th; The dialogue-based multiple-choice reading comprehension dataset DREAM BIBREF36 contains 10,197 questions for 6,444 multi-turn multi-party dialogues from English language exams that are designed by human experts to assess the comprehension level of Chinese learners of English. Compared with these datasets, ReClor distinguishes itself by targeting logical reasoning.

ReClor Data Collection and Analysis

ReClor Data Collection and Analysis ::: Data collection

The format of data in ReClor is similar to other multiple-choice reading comprehension datasets BIBREF10, BIBREF5, where a data point contains a context, a question and four answer options, among which only one option is right/most suitable. We collect reading comprehension problems that require complicated logical reasoning. However, producing such data requires the ability to perform complex logical reasoning, which makes it hard for crowdsourcing workers to generate such logical questions. Fortunately, we find the reading comprehension problems in some standardized tests, such as GMAT and LSAT, are highly in line with our expectation.

We construct a dataset containing 6,138 logical reasoning questions sourced from open websites and books. In the original problems, there are five answer options in which only one is right. To comply with fair use of law, we shuffle the order of answer options and randomly delete one of the wrong options for each data point, which results in four options with one right option and three wrong options. Furthermore, similar to ImageNet dataset, ReClor is available for non-commercial research purpose only. We are also hosting a public evaluation server on EvalAI BIBREF37 to benchmark progress on Reclor.

ReClor Data Collection and Analysis ::: Data analysis

As mentioned above, we collect 6,138 data points, in which 91.22% are from actual exams of GMAT and LSAT while others are from high-quality practice exams. They are divided into training set, validation set and testing set with 4,638, 500 and 1,000 data points respectively. The overall statistics of ReClor and comparison with other similar multiple-choice MRC datasets are summarized in Table TABREF9. As shown, ReClor is of comparable size and relatively large vocabulary size. Compared with RACE, the length of the context of ReCor is much shorter. In RACE, there are many redundant sentences in context to answer a question. However, in ReClor, every sentence in the context passages is important, which makes this dataset focus on evaluating the logical reasoning ability of models rather than the ability to extract relevant information from a long context. The length of answer options of ReClor is largest among these datasets. We analyze and manually annotate the types of questions on the testing set and group them into 17 categories, whose percentages and descriptions are shown in Table TABREF11. The percentages of different types of questions reflect those in the logical reasoning module of GMAT and LSAT. Some examples of different types of logical reasoning are listed in Figure FIGREF12, and more examples are listed in the Appendix . Taking two examples, we further express how humans would solve such questions in Table TABREF13, showing the challenge of ReClor.

ReClor Data Collection and Analysis ::: Data Biases in the Dataset

The dataset is collected from exams devised by experts in logical reasoning, which means it is annotated by humans and may introduce biases in the dataset. Recent studies have shown that models can utilize the biases in a dataset of natural language understanding to perform well on the task without truly understanding the text BIBREF13, BIBREF14, BIBREF15, BIBREF16, BIBREF17, BIBREF18. It is necessary to analyze such data biases to help evaluate models. In the ReClor dataset, the common context and question are shared across the four options for each data point, so we focus on the analysis of the difference in lexical choice and sentence length of the right and wrong options without contexts and questions. We first investigate the biases of lexical choice. We lowercase the options and then use WordPiece tokenization BIBREF39 of BERT$_{\small \textsc {BASE}}$ BIBREF1 to get the tokens. Similar to BIBREF16, for the tokens in options, we analyze their conditional probability of label $l \in \lbrace \mathrm {right, wrong}\rbrace $ given by the token t by $p(l|t) =count(t, l) / count(t)$. The larger the correlation score is for a particular token, the more likely it contributes to the prediction of related option. Table SECREF14 reports tokens in training set which occur at least twenty times with the highest scores since many of the tokens with the highest scores are of low frequency. We further analyze the lengths of right and wrong options BIBREF17 in training set. We notice a slight difference in the distribution of sentence length for right and wrong options. The average length for wrong options is around 21.82 whereas that for right options is generally longer with an average length of 23.06.

tableTop 10 tokens that correlate to right options with more than 20 occurrences. figureThe distribution of the option length in ReClor with respect to right and wrong labels.

Experiments ::: Baseline Models

Many neural network based models such as FastText BIBREF40, Bi-LSTM, GPT BIBREF41, GPT-2 BIBREF0, BERT BIBREF1, XLNet BIBREF2, RoBERTa BIBREF3 have achieved impressive results in various NLP tasks. We challenge these neural models with ReClor to investigate how well they can perform. Details of the baseline models and implementation are shown in the Appendix and .

Experiments ::: Experiments to Find Biased Data

As mentioned earlier, biases prevalently exist in human-annotated datasets BIBREF16, BIBREF17, BIBREF18, BIBREF42, which are often exploited by models to perform well without truly understanding the text. Therefore, it is necessary to find out the biased data points in ReClor in order to evaluate models in a more comprehensive manner BIBREF43. To this end, we feed the five strong baseline models (GPT, GPT-2, BERT$_{\small \textsc {BASE}}$, XLNet$_{\small \textsc {BASE}}$ and RoBERTa$_{\small \textsc {BASE}}$) with ONLY THE ANSWER OPTIONS for each problem. In other words, we purposely remove the context and question in the inputs. In this way, we are able to identify those problems that can be answered correctly by merely exploiting the biases in answer options without knowing the relevant context and question. However, the setting of this task is a multiple-choice question with 4 probable options, and even a chance baseline could have 25% probability to get it right. To eliminate the effect of random guess, we set four different random seeds for each model and pick the data points that are predicted correctly in all four cases to form the EASY set. Then, the data points which are predicted correctly by the models at random could be nearly eliminated, since any data point only has a probability of $(25\%)^{4}=0.39\%$ to be guessed right consecutively for four times. Then we unite the sets of data points that are consistently predicted right by each model, because intuitively different models may learn different biases of the dataset. The above process is formulated as the following expression,

where $_{\mathrm {BERT}}^{\mathrm {seed_1}}$ denotes the set of data points which are predicted correctly by BERT$_{\small \textsc {BASE}}$ with seed 1, and similarly for the rest. Table TABREF18 shows the average performance for each model trained with four different random seeds and the number of data points predicted correctly by all of them. Finally, we get 440 data points from the testing set $_{\mathrm {TEST}}$ and we denote this subset as EASY set $_{\mathrm {EASY}}$ and the other as HARD set $_{\mathrm {HARD}}$.

Experiments ::: Transfer learning Through Fine-tuning

Among multiple-choice reading comprehension or QA datasets from exams, although the size of ReClor is comparable to those of ARC BIBREF12 and DREAM BIBREF36, it is much smaller than RACE BIBREF5. Recent studies BIBREF44, BIBREF45, BIBREF25, BIBREF46 have shown the effectiveness of pre-training on similar tasks or datasets then fine-tuning on the target dataset for transfer learning. BIBREF46 find that by first training on RACE BIBREF5 and then further fine-tuning on the target dataset, the performances of BERT$_{\small \textsc {BASE}}$ on multiple-choice dataset MC500 BIBREF10 and DREAM BIBREF36 can significantly boost from 69.5% to 81.2%, and from 63.2% to 70.2%, respectively. However, they also find that the model cannot obtain significant improvement even performs worse if it is first fine-tuned on span-based dataset like SQuAD BIBREF4. ReClor is a multiple-choice dataset, so we choose RACE for fine-tuning study.

Experiments ::: Results and Analysis

The performance of all tested models on the ReClor is presented in Table TABREF21. This dataset is built on questions designed for students who apply for admission to graduate schools, thus we randomly choose 100 samples from the testing set and divide them into ten tests, which are distributed to ten different graduate students in a university. We take the average of their scores and present it as the baseline of graduate students. The data of ReClor are carefully chosen and modified from only high-quality questions from standardized graduate entrance exams. We set the ceiling performance to 100% since ambiguous questions are not included in the dataset.

The performance of fastText is better than random guess, showing that word correlation could be used to help improve performance to some extent. It is difficult for Bi-LSTM to converge on this dataset. Transformer-based pre-training models have relatively good performance, close to the performance of graduate students. However, we find that these models only perform well on EASY set with around 75% accuracy, showing these models have an outstanding ability to capture the biases of the dataset, but they perform poorly on HARD set with only around 30% accuracy. In contrast, humans can still keep good performance on HARD set. We notice the difference in testing accuracy performed by graduate students on EASY and HARD set, but this could be due to the small number of students participated in the experiments. Therefore, we say humans perform relatively consistent on both biased and non-biased dataset.

It is noticed that if the models are first trained on RACE and then fine-tuned on ReClor, they could obtain significant improvement, especially on HARD set. The overall performance of RoBERTa$_{\small \textsc{LARGE}}$ is even better than that of graduate students. This similar phenomenon can also be observed on DREAM dataset BIBREF36 by BIBREF46, which shows the potential of transfer learning for reasoning tasks. However, even after fine-tuning on RACE, the best performance of these strong baselines on HARD set is around 50%, still lower than that of graduate students and far away from ceiling performance.

Experiments in different input settings are also done. Compared with the input setting of answer options only (A), the setting of questions and answer options (Q, A) can not bring significant improvement. This may be because some questions e.g., Which one of the following is an assumption required by the argument?, Which one of the following, if true, most strengthens the argument? can be used in the same reasoning types of question, which could not offer much information. Further adding context causes significant boost, showing the high informativeness of the context.

We further analyze the model performance with respect to different question types of logical reasoning. Some results are shown in Figure FIGREF22 and the full results are shown in Figure , and in the Appendix . Three models of BERT$_{\small \textsc {LARGE}}$, XLNet$_{\small \textsc {LARGE}}$ and RoBERTa$_{\small \textsc {LARGE}}$ perform well on most of types. On HARD set, the three models perform poorly on certain types such as Strengthen, Weaken and Role which require extensive logical reasoning. However, they perform relatively better on other certain types, such as Conclusion/Main Point and Match Structures that are more straight-forward. For the result of transfer learning, we analyze XLNet$_{\small \textsc {LARGE}}$ in detail. Though the overall performance is significantly boosted after fine-tuning on RACE first, the histograms in the bottom of Figure FIGREF22 show that on EASY set, accuracy of the model with fine-tuning on RACE is similar to that without it among most question types, while on HARD set, significant improvement on some question types is observed, such as Conclusion/Main Point and Most Strongly Supported. This may be because these types require less logical reasoning to some extent compared with other types, and similar question types may also be found in RACE dataset. Thus, the pre-training on RACE helps enhance the ability of logical reasoning especially of relatively simple reasoning types, but more methods are still needed to further enhance the ability especially that of relatively complex reasoning types.

Conclusion

In this paper, we introduce ReClor, a reading comprehension dataset requiring logical reasoning, with the aim to push research progress on logical reasoning in NLP forward from sentence-level to passage-level and from simple logical reasoning to multiple complicated one. We propose to identify biased data points and split the testing set into EASY and HARD group for biased and non-biased data separately. We further empirically study the different behaviors of state-of-the-art models on these two testing sets, and find recent powerful transformer-based pre-trained language models have an excellent ability to exploit the biases in the dataset but have difficulty in understanding and reasoning given the non-biased data with low performance close to or slightly better than random guess. These results show there is a long way to equip deep learning models with real logical reasoning abilities. We hope this work would inspire more research in future to adopt similar split technique and evaluation scheme when reporting their model performance. We also show by first fine-tuning on a large-scale dataset RACE then fine-tuning on ReClor, the models could obtain significant improvement, showing the potential of transfer learning to solve reasoning tasks.

Conclusion ::: Acknowledgments

We would like to thank three anonymous reviews for their insightful comments and suggestions; thank Rishabh Jain from Georgia Tech for helping build up the leaderboard of ReClor on EvalAI. Jiashi Feng was partially supported by NUS IDS R-263-000-C67-646, ECRA R-263-000-C87-133, MOE Tier-II R-263-000-D17-112 and AI.SG R-263-000-D97-490.