LogiQA: A Challenge Dataset for Machine Reading Comprehension with Logical Reasoning

Jian Liu[1*], **Leyang Cui**[2,3*], **Hanmeng Liu**[2,3], **Dandan Huang**[2,3], **Yile Wang**[2,3], **Yue Zhang**[2,3†]
[1]School of Computer Science, Fudan University
[2]School of Engineering, Westlake University
[3]Institute of Advanced Technology, Westlake Institute for Advanced Study
jianliu17@fudan.edu.cn, {cuileyang, liuhanmeng, huangdandan, wangyile, zhangyue}@westlake.edu.cn,

Abstract

Machine reading is a fundamental task for testing the capability of natural language understanding, which is closely related to human cognition in many aspects. With the rising of deep learning techniques, algorithmic models rival human performances on simple QA, and thus increasingly challenging machine reading datasets have been proposed. Though various challenges such as evidence integration and commonsense knowledge have been integrated, one of the fundamental capabilities in human reading, namely logical reasoning, is not fully investigated. We build a comprehensive dataset, named **LogiQA**, which is sourced from expert-written questions for testing human **Logi**cal reasoning. It consists of 8,678 QA instances, covering multiple types of deductive reasoning. Results show that state-of-the-art neural models perform by far worse than human ceiling. Our dataset can also serve as a benchmark for reinvestigating logical AI under the deep learning NLP setting. The dataset is freely available at https://github.com/lgw863/LogiQA-dataset.

1 Introduction

Machine reading [Hermann *et al.*, 2015; Chen *et al.*, 2016] is a popular task in NLP, which is useful for downstream tasks such as open domain question answering and information retrieval. In a typical task setting, a system is given a passage and a question, and asked to select a most appropriate answer from a list of candidate answers. With recent advances of deep learning in NLP, reading comprehension research has seen rapid advances, with a development from simple factual question answering [Rajpurkar *et al.*, 2016] to questions that involve the integration of different pieces of evidences via multi-hop reasoning [Welbl *et al.*, 2018; Yang *et al.*, 2018] and questions that involve commonsense knowledge outside the given passage [Ostermann *et al.*, 2018; Huang *et al.*, 2019], where more varieties of challenges in human reading comprehension are investigated.

P1: David, Jack and Mark are colleagues in a company. David supervises Jack, and Jack supervises Mark. David gets more salary than Jack.

Q: *What can be inferred from the above statements?*
A. Jack gets more salary than Mark.
B. David gets the same salary as Mark.
C. One employee supervises another who gets more salary than himself.
✓ **D. One employee supervises another who gets less salary than himself.**

P2: Our factory has multiple dormitory areas and workshops. None of the employees who live in dormitory area A are textile workers. We conclude that some employees working in workshop B do not live in dormitory area A.

Q: *What may be the missing premise of the above argument?*
A. Some textile workers do not work in workshop B.
B. Some employees working in workshop B are not textile workers.
✓ **C. Some textile workers work in workshop B.**
D. Some employees living in dormitory area A work in the workshop B.

Figure 1: Examples of LogiQA. (✓ indicates the correct answer.)

One important aspect of human reading comprehension and question answering is logical reasoning, which was also one of the main research topics of early AI [McCarthy, 1989; Colmerauer and Roussel, 1996]. To this end, there has been relatively very few relevant datasets in modern NLP. Figure 1 gives two examples of such problems. In particular, P1 consists of a paragraph of facts, and a question that asks the testee to select a valid conclusion by taking the facts as premises. In order to select the correct candidate, a machine is expected to understand the premises and the candidate answers. The correct answer can be found by categorical reasoning. P2 of Figure 1 is more challenging in providing a premise and a conclusion in the paragraph, while asking for a missing premise. In particular, three sets of workers are involved, including those living in dormitory area A, those who work in workshop B and those who are textile workers. A testee can find the answer by drawing logical correlations between the three sets of workers.

The type of machine reading comprehension questions above requires a combination of natural language understanding and logical reasoning. Compared with factual question answering, lexical overlap between the paragraph and the candidate answers plays a relatively less important role. Compared with commonsense reading comprehension, such questions do not rely heavily on external knowledge. Instead they focus on logical inference. In this aspect, the questions can be viewed as a re-investigation of logical AI [McCarthy, 1989; Nilsson, 1991] in the age of neural NLP. One solution can be to perform semantic parsing into formal logic repre-

[*]Equal contribution.
[†]Corresponding Author

sentation, and then perform symbolic reasoning. However, with the abundance of deep learning methods in the NLP toolkit, it can also be interesting to learn the potentials of neural AI for solving such tasks.

To facilitate such research, we create a new reading comprehension dataset, LogiQA, which contains 8,678 paragraph-question pairs, each with four candidate answers. Our dataset is sourced from publically available logical examination papers for reading comprehension, which are designed by domain experts for evaluating the logical reasoning ability and test participants. Thus the quality and topic coverage of the questions are reliable. We manually select problems from the original dataset, filtering out problems that involve figures, charts, or those that are heavy of mathematics, and ensuring a wide coverage of logical reasoning types, including categorical reasoning, conditional reasoning, disjunctive reasoning and conjunctive reasoning [Hurley, 2014].

To establish baseline performances on LogiQA, we explore several state-of-the-art neural models developed for reading comprehension. Experimental results demonstrate a significant gap between machine (35.31% accuracy) and human ceiling performance (96.00%). We provide detailed analysis to give insights into potentially promising research directions.

2 Related Work

Existing Datasets For Reading Comprehension A seminal dataset for large-scale reading comprehension is SQuAD [Rajpurkar *et al.*, 2016], which requires selecting a factual answer from all possible spans in a given passage. Many neural methods have been developed for this dataset, achieving results that rival human testees. As a consequence, more reading comprehension datasets with increasing challenges are proposed. These datasets can be classified according to the main challenges. In particular, TriviaQA [Joshi *et al.*, 2017] requires evidence integration across multiple supporting documents to answer the questions. DuoRC [Saha *et al.*, 2018] and Narrative QA [Kočiský *et al.*, 2018] raise challenges by introducing two passages about the same facts. Welbl *et al.* [2018] and HotpotQA [Yang *et al.*, 2018] test models for text understanding with sequential multi-step reasoning. Drop [Dua *et al.*, 2019] tests discrete numerical reasoning over the context. MuTual [Cui *et al.*, 2020] tests dialogue reasoning ability via the next utterance prediction task. These datasets are factual in the sense that the answer (or candidate in multi-choice-questions) is mostly a text span in the given passage. Several types of reasoning are necessary, such as geolocational reasoning and numerical computation. Different from these datasets, our dataset contains answers not directly included in the input passage, and requires comprehensive reasoning methods beyond text matching based techniques.

Similar to our dataset, recent datasets for commonsense reasoning, including MCScript [Ostermann *et al.*, 2018] and COSMOS [Huang *et al.*, 2019], also contain candidate answers not directly included in the input passage. They test a model's capability of making use of external background knowledge about spatial relations, cause and effect, scientific facts and social conventions. In contrast, our dataset focuses on logical reasoning and most of the necessary facts are directly included in the given passage. In addition, most of the existing datasets are labeled by crowd sourcing. In contrast, our dataset is based on examination problems written by human experts for students, and therefore has a better guarantee of the quality. This is particularly important for datasets that involve abstract reasoning skills.

Dataset	Logic	Domain	Expert
SQuAD	✗	Wikipedia	✗
TriviaQA	✗	Trivia	✗
RACE	✗	Mid/High School Exams	✓
DuoRC	✗	Wikipedia & IMDb	✗
Narrative QA	✗	Movie Scripts, Literature	✗
DROP	✗	Wikipedia	✗
COSMOS	✗	Webblog	✗
MuTual	✗	Daily Dialogue	✓
LogiQA(Ours)	✓	Civil Servants Exams	✓

Table 1: Comparison with existing reading comprehension datasets. "Logic" indicates that dataset mainly requires logical reasoning. "Expert" indicates that dataset is designed by domain experts.

The correlation and difference between our dataset and existing QA datasets are shown in Table 1.

Logical Reasoning There have been existing datasets related to logical reasoning. In particular, Habernal *et al.*[2018] design a dataset for argument reasoning, where a claim is given and the model is asked to choose a correct premise from two candidates to support the claim. Similar to our dataset, the dataset concerns deductive reasoning. The biggest difference between our dataset and this dataset is that ours is a machine reading comprehension test while theirs focuses on argumentation. The form of their task is closer to NLI as compared with reading comprehension. In addition, our dataset has more instances (8,678 vs 1,970), more choices per question (4 vs 2) and is written by relevant experts rather than being crowd-sourced. CLUTRR [2019] is a dataset for inductive reasoning over family relations. The input is a given passage and a query pair, the output is a relationship between the pair. The dataset concerns reasoning on a fixed domain (i.e., family relationship), which is in line with prior work on social relation inference [Bramsen *et al.*, 2011]. In contrast, our dataset investigates general logical reasoning.

Datasets from Examinations Our dataset is also related to datasets extracted from examination papers, aiming at evaluating systems under the same conditions as how humans are evaluated. For example, the AI2 Elementary School Science Questions dataset [Khashabi *et al.*, 2016] contains 1,080 questions for students in elementary schools; RACE [Lai *et al.*, 2017] collects passages and questions from the English exams for middle school and high school Chinese students in ages between 12 to 18. These datasets are based on English tests, examining testees on general language understanding. They target students in language learning. In contrast, our datasets are based on examinations of logical skills in addition to language skills. To the best of our knowledge, LogiQA is the first large-scale dataset containing different types of logical problems, where problems are created based on exams designed by human experts.

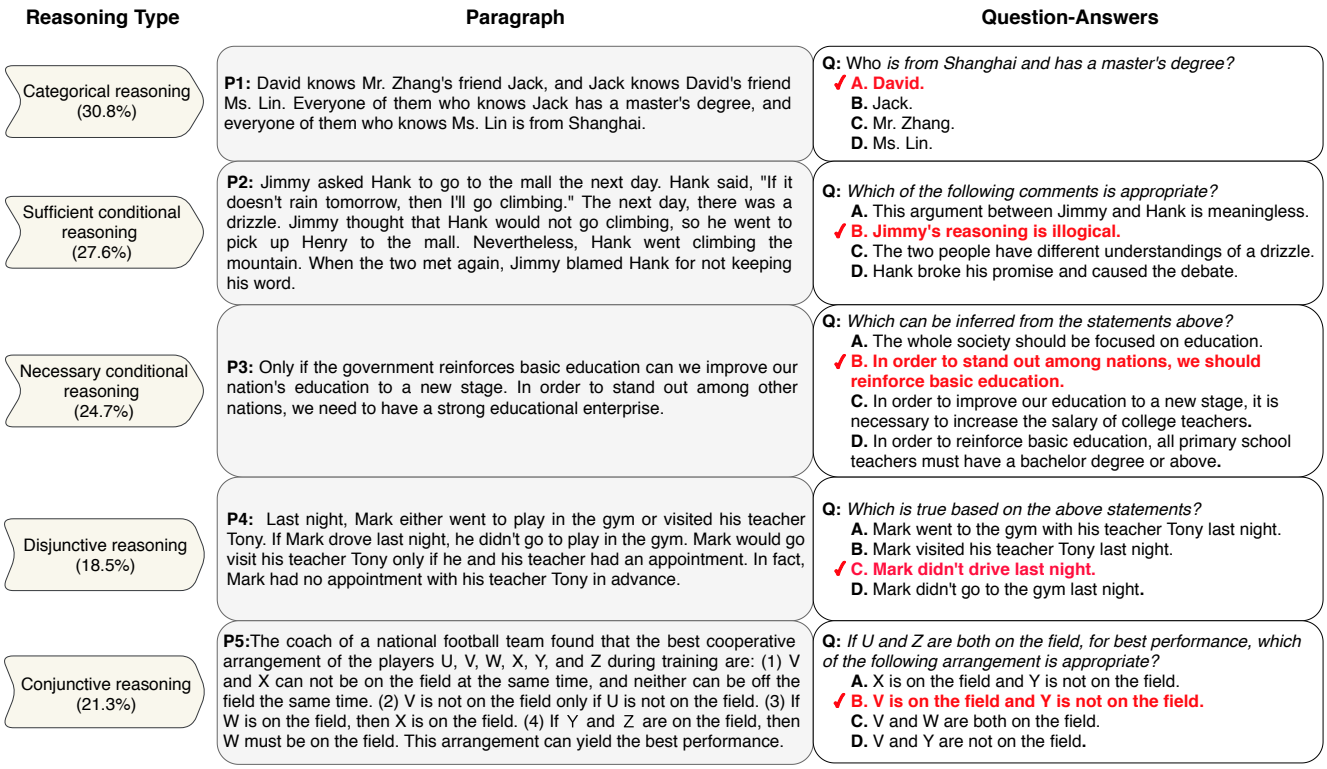

Figure 2: Examples of each type of logical reasoning in LogiQA. (✓ indicates the correct answer.)

3 Dataset

3.1 Data Collection and Statistics

We construct LogiQA by collecting the logical comprehension problems from publically available questions of the National Civil Servants Examination of China, which are designed to test the civil servant candidates' critical thinking and problem solving. We collected raw data released at the official website, obtaining 13,918 paragraph-question-choice triples with the correct answers.

The following steps are conducted to clean the raw data. First, we remove all the instances that do not have the format of our problem setting, i.e., a question is removed if the number of candidate choices is not four. Second, we filter all the paragraphs and questions that are not self-contained based on the text information, i.e. we remove the paragraphs and questions containing images or tables. We also remove all questions containing the keywords "underlined" and "sort sentences", since it can be difficult to reproduce the effect of underlines and sentence number order for a typical machine reader. Finally, we remove all duplicated paragraph-question pairs. The resulting dataset contains 8,678 paragraph-questions pairs.

Since the original dataset was written in Chinese, five professional English speakers are employed to translate the dataset manually. To ensure translation quality, we further employ three proofreaders. A translated instance is sent back to the translators for revision if proofreaders reject the instance. The detailed statistics for LogiQA is summarized in Table 2. Compared with existing reading comprehension datasets, the average paragraph length is relatively small since logical reasoning problems do not heavily rely on complex context.

Parameter	Value
# Paragraphs-Question Pair	8,678
Ave./Max. # Tokens / Paragraph	76.87 / 323
Ave./Max. # Tokens / Question	12.03 / 54
Ave./Max. # Tokens / Candidate Answer	15.83 / 111

Table 2: Statistics of LogiQA.

We also release the Chinese version of LogiQA (named as Chinese LogiQA) for Chinese reasoning-based reading comprehension research.

3.2 Reasoning Types of the Dataset

The test set of our benchmark consists of 867 paragraph-question pairs. We manually categorize the instances according to the five types of logical reasoning defined by Hurley [2014], including categorical reasoning, sufficient conditional reasoning, necessary conditional reasoning, disjunctive reasoning and conjunctive reasoning. These types of reasoning belong to deductive reasoning, for which a definite conclusion can be derived given a set of premises. As a result, such reasoning can be most suitable for evaluating performances quantitatively. Figure 2 shows the statistics and representative examples of the reasoning types in our dataset. Note that the sum of percentage values is above 100%, which is because one problem can involve multiple types of reasoning.

Formally, the five types of reasoning can be described as follows:

- **Categorical reasoning:** The goal is to reason whether a specific concept belongs to a particular category. This type of reasoning is commonly associated with quantifiers such as *"all/everyone/any"*, *"no"* and *"some"*, etc.
- **Sufficient conditional reasoning:** The type of hypothetical reasoning is based on conditional statements of the form "*If P, then Q*", in which *P* is the antecedent and *Q* is the consequent.
- **Necessary conditional reasoning:** This type of hypothetical reasoning is based on conditional statements of the form "*P* only if *Q*", "*Q* whenever *P*", etc., where Q is a necessary condition for P.
- **Disjunctive reasoning:** In this type of reasoning, the premises are disjunctive, in the form "*either . . . or . . .*" , where the conclusion holds as long as one premise holds.
- **Conjunctive reasoning:** In this type of reasoning, the premises are conjunctives, in the form "*both . . . and. . .* ", where the conclusion holds only if all the premises hold.

4 Methods

We evaluate the performances of typical reading comprehension models, including rule-based methods, deep learning methods as well as methods based on pre-trained contextualized embedding. In addition, human performances are evaluated and ceiling performances are reported.

Rule-Based Methods We adopt two rule-based methods, which rely on simple lexical matching. In particular, *word matching* [Yih *et al.*, 2013] is a baseline that selects the candidate answer that has the highest degree of unigram overlap with the given paragraph-question pair; *sliding window* [Richardson *et al.*, 2013] calculates the matching score for each candidate answer by extracting TF-IDF type features from n-grams in the given paragraph-question pair.

Deep Learning Methods Most existing methods [Chen *et al.*, 2016; Dhingra *et al.*, 2017; Wang *et al.*, 2018] find the answer by text matching techniques, calculating the similarity between the given paragraph, the question and each candidate answer. In particular, *Stanford attentive reader* [Chen *et al.*, 2016] computes the similarity between a paragraph-question pair and a candidate answer using LSTM encoding and a bi-linear attention function; *gated attention reader* [Dhingra *et al.*, 2017] adopts a multi-hop architecture with a more fine-grained mechanism for matching candidate answers with paragraph-question pairs; *co-matching network* [Wang *et al.*, 2018] further enhances matching the paragraph-question pair and paragraph-candidate answer pair by encoding each piece of text and calculating matching score between each pair, respectively.

Pre-trained Methods Pre-trained models give the current state-of-the-art results on machine reading. Different from the above deep learning methods, pre-trained methods consider the paragraph, the question and each candidate answer as one concatenated sentence, using a pre-trained contextualized embedding model to encode the sentence for calculating its score. Given four candidate answers, four concatenated sentences are constructed by pairing each candidate answer with the paragraph and question, and the one with the highest model score is chosen as the answer. In particular, *BERT* [Devlin *et al.*, 2019] treats the paragraph as sentence A and the concatenation of the question and each candidate as sentence B, before further concatenating them into *[CLS] A [SEP] B [SEP]* for encoding; *RoBERTa* [Liu *et al.*, 2019] replaces the BERT model using the RoBERTa model. The hidden state of the *[CLS]* token is used for MLP + softmax scoring. The embedding models are fine-tuned during training.

Human Performance We employ three post-graduate students for human performance evaluation, reporting the average scores on 500 randomly selected instances from the test set. For calculating the ceiling performances, we consider a question as being correctly answered if one of the students gives the correct answer.

Implementation Details We re-implement the rule-based methods strictly following the original papers [Yih *et al.*, 2013; Richardson *et al.*, 2013]. For the deep learning methods, we directly use the implementations released in the original papers. 100-dimensional Glove word embeddings are used as embedding initialization. For pre-trained methods, we follow the HuggingFace implementation [Wolf *et al.*, 2019]. We take the off-the-shelf model BERT-base and RoBERTa-base for LogiQA, and Chinese BERT-base and Chinese RoBERTa-base [Cui *et al.*, 2019] for Chinese LogiQA. All hyper-parameters are decided by the model performance on the development sets.

5 Results

We randomly split the dataset, using 80% for training, 10% for development and the remaining 10% for testing. Table 3 shows the results of the models discussed in the previous section. In particular, the human performance is 86.00% and the ceiling performance is 95.00%, which shows that the difficulty level of the dataset is not high for human testees. In contrast, all of the algorithmic models perform significantly worse than human, demonstrating that the methods are relatively weak in logical reasoning reading comprehension. In addition, results on the Chinese dataset are on the same level compared with those on the English dataset.

In particular, the rule-based methods give accuracies of 28.37% and 22.51%, respectively, the latter being even lower than a random guess baseline. This shows that the questions are extremely difficult to solve by using lexical matching alone. Figure 1 serves as one intuitive example. The deep learning methods such as the Stanford attentive reader, the gated attention reader and the co-matching network give accuracies around 30%, which is better compared with the random guess baseline but far behind human performance. One likely reason is that the methods are trained end-to-end, where it turns out difficult for attention-based text matching to learn underlying logical reasoning rules.

It has been shown that pre-trained models have a certain degree of commonsense and logical capabilities [Huang *et al.*,

Category	Model	LogiQA		Chinese LogiQA	
		Dev	**Test**	**Dev**	**Test**
	Random(theoretical)	25.00	25.00	25.00	25.00
Rule-based	Word Matching [Yih *et al.*, 2013]	27.49	28.37	26.55	25.74
	Sliding Window [Richardson *et al.*, 2013]	23.58	22.51	23.85	24.27
Deep learning	Stanford Attentive Reader [Chen *et al.*, 2016]	29.65	28.76	28.71	26.95
	Gated-Attention Reader [Dhingra *et al.*, 2017]	28.30	28.98	26.82	26.43
	Co-Matching Network [Wang *et al.*, 2018]	33.90	31.10	30.59	31.27
Pre-trained	BERT [Devlin *et al.*, 2019]	33.83	32.08	30.46	34.77
	RoBERTa [Liu *et al.*, 2019]	**35.85**	**35.31**	**39.22**	**37.33**
Human	Human Performance	-	86.00	-	88.00
	Ceiling Performance	-	95.00	-	96.00

Table 3: Main results on LogiQA (accuracy%).

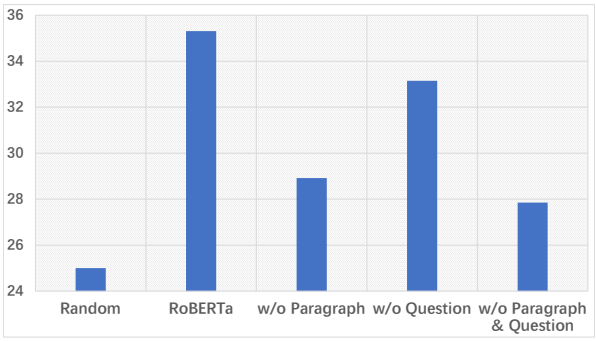

Figure 3: Ablation of paragraph or question (accuracy%).

Model	Dev	Test
Random(theoretical)	25.00	25.00
$\text{RoBERTa}_{\text{LogiQA}}$	35.85	35.31
$\text{RoBERTa}_{\text{RACE}}$	29.19	26.86
$\text{RoBERTa}_{\text{COSMOS}}$	25.14	28.73
$\text{RoBERTa}_{\text{RACE}\longrightarrow\text{LogiQA}}$	34.60	35.07
$\text{RoBERTa}_{\text{COSMOS}\longrightarrow\text{LogiQA}}$	36.44	35.11

Table 4: Transfer learning results (accuracy%).

Length	(0,100]	(100,150]	(150,200]	(200,+∞)
#Instances	253	364	198	61
RoBERTa	31.31	36.25	37.13	40.38

Table 5: Performance of different length (accuracy%).

2019]. On LogiQA, such models give better performances compared with the methods without contextualized embeddings. However, the best result by RoBERTa is 35.31%, still much below human performance. This shows that knowledge in pre-trained models is rather weak for logical reasoning. It remains an open question on how deep learning machine readers can be equipped with strong reasoning capability.

6 Discussion

We give detailed analysis based on the empirical results of RoBERTa and other models on LogiQA test set.

6.1 Ablation Tests

Following recent studies, we conduct a set of ablation experiments using RoBERTa to measure bias in the dataset by checking the performance based on the partial information [Cai *et al.*, 2017]. Figure 3 shows the results on the test set. There is a significant drop of accuracies without the paragraph, the question or both, which indicates that the bias on the dataset is weak. In particular, without the input paragraph, the accuracy drops from 35.31% to 28.92%, which is comparable to a drop from 67.1% to 55.9% by the same model on the COSMOS dataset [Huang *et al.*, 2019], and 66.0% to 52.7% on the Social IQa dataset [Sap *et al.*, 2019].

Ablating question causes a relatively smaller performance drop as compared with the paragraph, which is consistent with observations by Huang *et al.*[2019]. This is likely because the diversity of questions is lower. The above results show that our dataset does not have a strong bias.

6.2 Transfer Learning

We conduct a set of transfer learning experiments to understand the degree of overlap in terms of necessary knowledge for solving problems in our dataset and existing datasets. In particular, we first fine-tune the RoBERTa model on a source dataset, before fine-tuning the model on LogiQA. If the required knowledge is similar, the model performance is expected to increase. RACE and COSMOS are adopted as the source datasets. The former tests English reading skills while the latter tests commonsense knowledge. As shown in Table 4, the RoBERTa model trained only on either source dataset gives significantly lower accuracies on LogiQA test set compared with the RoBERTa model trained on LogiQA. The performance of RoBERTa trained on RACE is even close to the random guess baseline. In addition, further fine-tuning on LogiQA leads to improvements over the source-trained baselines, but the resulting models do not outperform a model trained only on LogiQA. The observation is different from most other datasets [Huang *et al.*, 2019; Sap *et al.*, 2019], which demonstrates that LogiQA contains highly different challenges compared with existing datasets.

6.3 Performance Across Different Lengths

We measure the accuracy of RoBERTa against the input size. In particular, the number of words in the paragraph, the question and the candidate answers are added together as the length of a test instance. The statistics and performances are all shown in Table 5. Interestingly, the model performances are not negatively associated with the input size, which is different from most NLP benchmarks. This shows that the level of challenge in logical reasoning can be independent of the input verbosity.

P1: Children's products are any products intended for play or use by children 12 years of age or younger.

Q: *Based on the above definition, which of the following are children's products?*
A. Milk powders for infants aged from 0 to 1.
✓ **B. Comic books suitable for kids around 10 years old.**
C. Brightly packed lollipops.
✗ **D. Bumper cars in the amusement park children love to play.**

P2: Flower Bay is an ideal river for salmons swimming. If there is a hydropower dam downstream, then salmons will not be able to swim here. Salmons swim here only if the trees on the shore of Flower Bay have lost their leaves. If many sea eagles and brown bears gather in this river bay, then you can tell that the salmons are migrating. Now there are a lot of salmons swimming in Flower Bay.

Q: *Based on the above statements, which of the following can be derived?*
✓ **A. The leaves on the shore of Flower Bay are gone.**
B. There are many sea eagles and brown bears in Flower Bay.
✗ **C. There is a hydropower dam downstream of Flower Bay.**
D. Sea Eagle and Brown Bear Feed on Salmon.

P3: A company decided to select 4 people from 3 women (A, B, C) and 5 men (D, E, F, X, Y) to set up a group for an important negotiation. Here are the prerequisites: (1) The group members must have both women and men. (2) D and A cannot be selected at the same time. (3) B and C cannot be selected at the same time. (4) If Y is selected, then F won't be selected.

Q: *If D must be selected, which of the following can be derived?*
A. If the company selects F, then need also select Y.
✗ **B. If the company selects E, then need also select X.**
C. Either selecting Y or X.
✓ **D. Either selecting B or C.**

Figure 4: Example mistakes of RoBEETa. (✓ indicates the correct answers and ✗ indicates the RoBERTa prediction.)

6.4 Lexical Overlap

We aim to understand a bias of models in selecting the candidate answers that have the best surface matching with the paragraph. To this end, we calculate the unigram overlap between each candidate answer and the given paragraph for each problem, and mark the best-matching candidate. We report the "Overlap Ratio" by calculating the accuracy between model prediction and the best-matching candidate. The results are shown in Table 6. As can be seen, the gold-standard output has an accuracy of 28.37%, whilst all of the models give accuracies above this number, which shows a tendency of superficial matching. In particular, the word matching method gives an accuracy of 100% due to its mechanism. RoBERTa gives the lowest matching accuracy, showing that it relies the least on lexical patterns. We additionally measure the accuracy of the models on the "correct" instances according to best matching. RoBERTa still outperforms the other models, demonstrating relative strength in logical reasoning.

6.5 Reasoning Types

Table 7 gives the performances of RoBERTa over the 5 reasoning types discussed in Section 3.2. The method gives the best accuracy on categorical reasoning. However, for the other four reasoning types, the results are significantly lower. To understand why these tasks are challenging for RoBERTa, we give qualitative discussion via case study.

Categorical reasoning: P1 of Figure 4 shows a typical example, where the definition of children's products is given in the paragraph and the testee is asked to select a correct instance. A key here is the age range (i.e., under 12). RoBERTa incorrectly chooses the candidate that is superficially similar to the paragraph, while ignoring the reasoning process.

Model	Overlap Ratio	Accuracy(%)
Word Matching	100.00	28.37
Stanford Attentive Reader	35.47	35.82
Gated-Attention Reader	37.33	34.96
Co-Matching Network	40.74	36.85
BERT	34.23	37.73
RoBERTa	32.08	40.38
Gold-standard	28.37	100.00

Table 6: Overlap ratio (%) against the model type.

Reasoning Type	RoBERTa
Categorical reasoning	55.00
Sufficient conditional reasoning	17.11
Necessary conditional reasoning	19.29
Disjunctive reasoning	22.67
Conjunctive reasoning	21.98

Table 7: RoBERTa accuracy (%) against the reasoning type.

Conditional reasoning: P2 of Figure 4 is a representative example of the most challenging conditional reasoning questions. In particular, a variety of sufficient and necessary conditional relations are given in the paragraph, which include:

$x =$ "Salmons swim"
$y =$ "Sea eagles and brown bears gather"
$z =$ "Hydropower dam exists downstream"
$w =$ "Trees lose leaves"
$x => w$ (Necessary conditional relation)
$y => x$ (Sufficient conditional relation)
$x => \bar{z}$ (Sufficient conditional relation)

The correct answer depends on fully understanding both the necessary and sufficient conditional reasoning facts. RoBERTa makes a mistake by ignoring the "*not*" operator in the $x => \bar{z}$ condition, which coincides with prior observations on BERT and negation [Niven and Kao, 2019].

Conjunctive and disjunctive reasoning: P3 of Figure 4 represents one of the most challenging questions in the dataset, where the premises and candidate give a set of constraints in both conjunctive and disjunctive forms, and the question asks which candidate conforms to the premises. The testee is expected to enumerate different possible situations and then match the cases to the candidates by thoroughly understanding the candidates also. Intuitively, RoBERTa is not directly equipped with such reasoning capacity.

7 Conclusion

We have presented LogiQA, a large-scale logical reasoning reading comprehension dataset. In addition to testing reasoning capacities of machine reading, our dataset can also serve as a benchmark for re-examining the long pursued research of logical AI in the deep learning NLP era. Results show that the state-of-the-art machine readers still fall far behind human performance, making our dataset one of the most challenging test for reading comprehension.

Acknowledgments

This work is supported by the National Science Foundation of China (Grant No. 61976180). We also acknowledge funding support from the Westlake University and Bright Dream Joint Institute for Intelligent Robotics.

References

[Bramsen *et al.*, 2011] P. Bramsen, M. Escobar-Molano, A. Patel, and R. Alonso. Extracting social power relationships from natural language. In *NAACL*, 2011.

[Cai *et al.*, 2017] Z. Cai, L. Tu, and K. Gimpel. Pay attention to the ending: Strong neural baselines for the roc story cloze task. In *ACL*, 2017.

[Chen *et al.*, 2016] D. Chen, J. Bolton, and C. D. Manning. A thorough examination of the CNN/daily mail reading comprehension task. In *ACL*, 2016.

[Colmerauer and Roussel, 1996] A. Colmerauer and P. Roussel. The birth of prolog. In *History of programming languages—II*. ACM, 1996.

[Cui *et al.*, 2019] Y. Cui, W. Che, T. Liu, B. Qin, Z. Yang, S. Wang, and G. Hu. Pre-training with whole word masking for chinese bert. *arXiv*, 2019.

[Cui *et al.*, 2020] L. Cui, Y. Wu, S. Liu, Y. Zhang, and M. Zhou. Mutual: A dataset for multi-turn dialogue reasoning. In *ACL*, 2020.

[Devlin *et al.*, 2019] J. Devlin, M-W Chang, K. Lee, and K. Toutanova. BERT: Pre-training of bidirectional transformers for language understanding. In *NAACL*, 2019.

[Dhingra *et al.*, 2017] B. Dhingra, H. Liu, Z. Yang, W. Cohen, and R. Salakhutdinov. Gated-attention readers for text comprehension. *ACL*, 2017.

[Dua *et al.*, 2019] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *NAACL*, 2019.

[Habernal *et al.*, 2018] I. Habernal, H. Wachsmuth, I. Gurevych, and B. Stein. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *NAACL*, 2018.

[Hermann *et al.*, 2015] K. Moritz Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. In *NIPS*, 2015.

[Huang *et al.*, 2019] L. Huang, R. Le Bras, C. Bhagavatula, and Y. Choi. Cosmos QA: Machine comprehension with contextual commonsense reasoning. In *EMNLP*, 2019.

[Hurley, 2014] Patrick J. Hurley. *A concise introduction to logic*. Nelson Education, 2014.

[Joshi *et al.*, 2017] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*, 2017.

[Khashabi *et al.*, 2016] D. Khashabi, T. Khot, A. Sabharwal, P. Clark, O. Etzioni, and D. Roth. Question answering via integer programming over semi-structured knowledge. *arXiv*, 2016.

[Kočiský *et al.*, 2018] Tomáš Kočiský, J. Schwarz, P. Blunsom, C. Dyer, K. M. Hermann, Gábor Melis, and E. Grefenstette. The NarrativeQA reading comprehension challenge. *TACL*, 2018.

[Lai *et al.*, 2017] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy. RACE: Large-scale Reading Comprehension dataset from Examinations. In *EMNLP*, 2017.

[Liu *et al.*, 2019] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv*, 2019.

[McCarthy, 1989] John McCarthy. Artificial intelligence, logic and formalizing common sense. In *Philosophical logic and artificial intelligence*. 1989.

[Nilsson, 1991] Nils J. Nilsson. Logic and artificial intelligence. *Artificial intelligence*, 1991.

[Niven and Kao, 2019] T. Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. In *ACL*, 2019.

[Ostermann *et al.*, 2018] S. Ostermann, M. Roth, A. Modi, S. Thater, and M. Pinkal. Task 11: Machine comprehension using commonsense knowledge. In *SemEval*, 2018.

[Rajpurkar *et al.*, 2016] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *EMNLP*, 2016.

[Richardson *et al.*, 2013] M. Richardson, C.J.C. Burges, and E. Renshaw. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *EMNLP*, 2013.

[Saha *et al.*, 2018] A. Saha, R. Aralikatte, Mitesh M. Khapra, and K. Sankaranarayanan. DuoRC: Towards complex language understanding with paraphrased reading comprehension. In *ACL*, 2018.

[Sap *et al.*, 2019] M. Sap, H. Rashkin, D. Chen, R. Le Bras, and Y. Choi. Social iqa: Commonsense reasoning about social interactions. *EMNLP*, 2019.

[Sinha *et al.*, 2019] K. Sinha, S. Sodhani, J. Dong, J. Pineau, and W. L. Hamilton. Clutrr: A diagnostic benchmark for inductive reasoning from text. In *EMNLP*, 2019.

[Wang *et al.*, 2018] S. Wang, M. Yu, J. Jiang, and S. Chang. A co-matching model for multi-choice reading comprehension. In *ACL*, 2018.

[Welbl *et al.*, 2018] J. Welbl, P. Stenetorp, and S. Riedel. Constructing datasets for multi-hop reading comprehension across documents. *TACL*, 2018.

[Wolf *et al.*, 2019] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, Rémi Louf, M. Funtowicz, and J. Brew. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, 2019.

[Yang *et al.*, 2018] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*, 2018.

[Yih *et al.*, 2013] Wen-tau Yih, Ming-Wei Chang, C. Meek, and A. Pastusiak. Question answering using enhanced lexical semantic models. In *ACL*, 2013.