Summary

In Machine Reading Comprehension (MRC), a computational model is given some context and a question related to the context, with the goal of outputting an answer to the question. An example of this could be the following context-question-answer triple:

- Context: *"Polar bears live in the Arctic."*
- Question: *"Where do polar bears live?"*
- Answer: *"The Arctic."*

MRC has a problem though. Models have been outperforming baseline human performances on many benchmark datasets, such as the highly popular SQuAD and SuperGLUE. However, these model benchmark performances are not indicative of their realistic capabilities. Studies have shown that when the same models are given another benchmark that is similar, where one should expect a score that is just as good as the first benchmark, the models seem to have a drastic fall in performance. An example is how BERT (Devlin et al., 2018) can not handle simple changes in sentence structure that preserves the meaning of the text. This problems has led researchers to analyse how models figure out their answer, and it appears to be through training set memorization, heuristics, and statistical guessing (McCoy et al., 2019; Lewis et al., 2020).

The goal then is to push models towards a different approach that relies more on comprehension. We have decided to analyse datasets across five criteria for the purpose of measuring their suitability for MRC. These criteria are: bias, size, grammar, validity, and robustness. Bias is both a social and computational problem. High amounts of bias in a training set can lead to low performance on benchmarks, simply because the model is biased to answer in a particular way. We measure this by reading into what consideration the creators of a dataset have done to prevent bias from forming in their dataset. It seems impossible to solve all bias problems, and we simply encourage the community to be mindful when creating datasets. The size is important, because a model needs to learn through a lot of examples in order to produce a suitable output. Heuristics seem to suggest that 100k questions for a dataset is an appropriate amount. This is not a definitive cut-off between good and bad datasets, because difficult datasets have different standards. Grammar has been shown to increase performance, especially proper spelling. Checking the spelling of a piece of text is easy. Simply check all words against a dictionary, and then filter out names. What is left should be considered grammatical errors. Validity refers to making sure that nothing is vague, everything in a sentence makes sense, and the text is consistent. This is difficult to check algorithmically, and we therefore propose ensuring this criterion during the creation of the dataset. Datasets are mostly created through manual annotations by crowdworkers. These workers would then have to peer-review other's work to ensure everyone understands the piece of text in the same way. The biggest problem is the lack of robustness in many datasets. Robustness builds upon the foundation that generalizability should be considered a more desirable trait in MRC models. Robustness assumes that it is impossible to know which kinds of text a model will be presented with, and therefore it is important to include as many different kinds as possible in the training set. This includes dimensions such as length, difficulty, and domain knowledge. Domain knowledge is seemingly impossible to accurately measure. This is because a perfect measurement would require knowing every single fact within the domain that the model is expected to operate in, and if we had that, then the problem was solved from the start. For the purpose of this paper, we limited robustness to only account for the length of the given texts.

We set up experiments to that uses these methods to measure the presence of our criteria in a given dataset. We performed this experiment on 31 datasets in total. These datasets were chosen by popularity on the website `paperswithcode.com`. If we account for all these criteria, it is difficult to say that current state-of-the-art datasets really fulfill their purpose of MRC. When looking at Table 6, we can see that not many datasets fulfill more than maybe one or two criteria. The most outstanding dataset is by far SuperGLUE, being one of the best in all criteria. SQuAD is an extremely popular dataset, but it does not do well in our criteria. Something like ReCoRD is highly robust and with good verification, but it is contained in SuperGLUE.

We present a detailed explanation of how one can make sure their dataset creation process adheres to our criteria. We go into detail with the different phases of creating a dataset while also describing how we would build SQuAD 2.0 based on our criteria versus how the dataset was built. We break down the dataset and discuss the consideration of how we would integrate the criteria into SQuAD 2.0.

Including comprehension in Machine Reading Comprehension

Alexander Lenni Lausten Christian Ølund Dovydas Jasulaitis Karl Immanuel Holt

{alaust19, calund19, djasul19, kholt19}@student.aau.dk

Abstract

Machine Reading Comprehension (MRC) has a comprehension problem. Studies have shown that current state-of-the-art models are not able to perform realistically as well as they have done on benchmarks. This is due to various factors that imitate the concept of text comprehension but are not as flexible in their approach. In this paper, we argue how and why training datasets, and by extension computational models, are insufficient for the task of comprehension. We present five detailed criteria, which generally apply to any MRC dataset, and analyse to which extent current state-of-the-art datasets fulfill these criteria. These criteria are disincentivizing biases, adequate statistical power, grammar checking, verification, and robustness. We observe that many of the popular datasets fall short when being measured against these criteria. Moreover, we discuss what could have been improved in the creation of the popular SQuAD 2.0 dataset and how we would have ensured our own criteria in this process. We conclude that there is room for improvement in the overall creation of datasets, and we emphasize a greater focus, especially on biases and robustness.

1 Introduction

In Machine Reading Comprehension (MRC), a computational model is given some context and a question related to the context, with the goal of outputting an answer to the question. An example of this could be the following context-question-answer triple:

- Context: *"Polar bears live in the Arctic."*
- Question: *"Where do polar bears live?"*
- Answer: *"The Arctic."*

Computational models that can perform MRC are expected to solve tasks within a single or multiple domains by using MRC skills. A task is a single instance of the context-question-answer triple. A variety of domains, or fields that the model is expected to perform tasks within, have been studied, such as medical science, general QA, or Google News Articles. A skill is the proficiency of a type of language structure required to either understand the context or question, or produce an answer. Examples of these skills include multi-hop reasoning, coreference resolution and common sense reasoning. Many of the highly popular datasets have been solved, meaning they have benchmarks that exceed the baseline human performance. GLUE has a human performance of 87.6% while the highest performing model has 91.3%, which means the models outperform humans by an absolute value of 3.7%. Models have also exceeded human performance by an absolute value of 1.4% on SuperGLUE. The popular SQuAD 2.0 dataset has a human performance of 86.8%, while the highest scoring model has a performance of 90.9%, resulting in an absolute value of 4.1% over human performance. These benchmarks are inaccurate, as these highly performing models will not perform as expected when tested in a zero-shot environment. This is a problem that seemingly permeates a few other subfields in Natural Language Processing (NLP), such as Natural Language Inference and Neural Machine Translation (Schlegel et al., 2020; Zhang et al., 2020). Many researchers responsible for current state-of-the-art models have taken detrimental shortcuts in pursuit of increasing their model's ability to score highly in some metrics. This has left the concept of comprehension disregarded as a metric for how the model actually figures out the candidate answer. McCoy et al. (2019) as well as Lewis et al. (2020) argue that many state-of-the-art models rely on statistical patterns, training set memorization and heuristics that will succeed in the majority of test examples rather than learning the underlying comprehension they are intended to capture. Compounding this problem is the fact that little concern is afforded to the situation in general. Much of the current literature indicates it is more popular to ignore this in favor of squeezing out better numbers on a leaderboard (Church, 2017). Some studies even show that professionals in the field are free to optimize the testing sets to perfectly suit the model, thus inflating the scores (Bowman and Dahl, 2021). These factors together mean that testing a current state-of-the-art model on a dataset that should be relevant for the model can yield poor results, because the initial training dataset was not adequate for the task of comprehending the given text. Ribeiro et al. (2020) shows a detailed breakdown of situations where BERT (Devlin et al., 2018) will give the wrong answer, such as by changing sentence structures, using spatiotemporal concepts, or exposing BERT's bias. These challenges should be within the expected domain of BERT, but as an example the model has a 100% failure rate on any question containing coreferencing.

We propose five criteria that will serve as a measure of a datasets suitability for MRC. We have collected 31 datasets, which we analyse based on our criteria. We observe that SQuAD is overrepresented in relation to its performance on our criteria. We also observe that across multiple criteria, SuperGLUE is an extremely effective dataset due to its broad variety of tasks and composites. The worst dataset we found was WikiQA, with a high level of grammatical errors and ambiguity. We note that by far the most usual approach to making a dataset is to webscrape some context. This is then given to crowdworkers that will annotate a question and an answer based on the given context.

We then lay out a blueprint for how we ideally would create a dataset that fulfills our own criteria and how other dataset creators can approach making a better dataset in the future. We will also discuss the creation of SQuAD 2.0 and how we would approach the process in order to ensure our own criteria. We will not consider the computational efficiency required to process a dataset that fulfills our criteria.

Defining comprehension

Models are not currently capable of proving they actually understand their given text, but rather rely on statistical guessing, training set memorization, and heuristics that just happen to be often correct in their domains and skills. Psychologists have tried to argue what reading comprehension actually means for humans, and how computers can get closer to it. Sugawara et al. (2020) discusses a lot of dimensions through which a text can be understood, such as “reading between the lines” and understanding implicit motivations and causes that are never directly described in the text. One aspect they emphasize is the generalizability of the model and dataset. Generalizability is the opposite of specialization. Where specialization emphasizes efficiency at tasks within a domain, generalizability means being good at a bigger variety of tasks, within or outside a domain. This would usually imply a loss of performance in the generalized model, but Talmor and Berant (2019) has shown that generalized models perform close to or better than state-of-the-art models in their given domain, while also extremely outperforming the specialized models in other domains. However, this concept of more generalizability always being better presents a problem. It is a highly challenging task to determine when a dataset fully encompasses its required domain, and the model is able to generalize as needed for the purpose of its task. More generalizability also means more data of a broader variety, and this is an important factor that increases the cost of making a dataset. Some of the current literature argues that making optimal datasets is often constrained by funding (Bowman and Dahl, 2021). Others have circumvented this problem by simply training on multiple datasets that already exist.

The most important thing is to move models towards methods of comprehension that are more adaptable, especially in the domain a model is supposed to operate.

2 Criteria for good datasets

In this section, we present criteria for creating a dataset, along with how to measure and validate those criteria. These criteria are created to address the current issues in MRC (Bowman and Dahl, 2021; Dzendzik et al., 2021), such as overfitting and training set memorization. After a thorough analysis on MRC and datasets, we propose five criteria that a dataset should fulfill in order to be considered adequate for reading comprehension. These criteria have multiple dimensions through which one can measure their presence in a dataset. If one were to accurately measure all of these dimensions, one would be able to quantify the suitability of a dataset for the task of comprehension. We will propose some ways to measure these dimensions, albeit some of them appear to be very difficult and would require a much larger scope.

2.1 Disincentivizing biases

The importance of the problem of disincentivizing biases has been overlooked or underplayed by a number of current state-of-the-art datasets, as will be discussed in this work. Just as Markl and McNulty (2022) emphasize that it is dangerous to disregard biases in Automatic Speech Recognition (ASR), it is as dangerous in MRC, since MRC is just as likely to be used as a building block in a bigger system, which might carry on these biases. Systems can often better their chances on a dataset by adopting heuristics that reproduce the potential-harmful biases, which occurs in the dataset, often built by crowdworkers or naturally-occurring texts (Bowman and Dahl, 2021). Rudinger et al. (2017) and Bolukbasi et al. (2016) show how relying on crowdworkers and naturally occurring text are inherently biased. In (Rudinger et al., 2017), they additionally find biases for racial, religious and age-based stereotypes in the Stanford Natural Language Inference - one of the biggest datasets.

(Markl and McNulty, 2022) discuss how some dialects might unintentionally not be supported in ASR, simply because of a lack of representation. In the same way, some words, spellings or phrases used by minorities, might not be supported or handled unintentionally, in MRC. All of these small changes is exactly what combined as a culture, showcasing how important it is to take it into consideration when relying on a simple dictionary (Frisch, 1964).

Bolukbasi et al. (2016) are the first to ask questions related to biases in word embedding, a method which has shown great results, and is proposed for a variety of problems. They show that they are able to bring gender biases down to a third using their model, and prove there is a lot of space for research on the topic of disincentivizing biases. It is, however, difficult to create a metric to look for biases against minorities, religions, genders, and so on. Developing such a metric is no easy task and should not be taken lightly, since these are groups which often change, and overlooking certain groups, either intentionally or by accident, is a political statement (Bowman and Dahl, 2021). The difficulty in developing the metric and the ever changing targeted groups is not a valid argument to ignore looking out for known biases. Bowman and Dahl (2021) proposes to benchmark a model on a set of anti-bias datasets in parallel with the benchmarking on the primary dataset. These kind of anti-bias benchmarks are already available in many fields. A further problem is to make the researchers publish these results along with their paper, since disclosing small potential problems on bias benchmarks is, in the current research environment, worse than disclosing nothing. Bowman and Dahl (2021) proposes the solution to be a better research culture, where to get through the peer-review process, these data should be published.

Markl and McNulty (2022) warn that it might be the market that controls the direction of research, and therefore not give any focus to disincentivizing biases if it is not profitable. Therefore, it is much more important that universities demand a form for testing for biases and a publicity of these results. Markl and McNulty (2022) propose a civic design as the solution for ASR being controlled by the market, civic design allowing the direction and detection of biases being controlled by nations or maybe by organisations like the UN, instead of the market.

In this paper, we aim to be able to measure the degree by which a given dataset disincentivises biases. To this end, a scoring system is presented in Tables 1 and 2, with a combined maximum of 3 points and a minimum of 0, thus higher the score the better. The two Tables 1 and 2 presents scoring descriptions to measure the papers attention to biases in their dataset and the papers effort to make the community aware of potential biases and encouraging to prevent biases respectively. In this paper, attention to biases is evaluated at twice as many points as efforts. This is because we see it as a fundamental criteria for the dataset creators to think abut which biases they are opening the community up to, by publishing a biased dataset. Effort is more about trying to catch biases, after the fact.

2.2 Adequate statistical power

Adequate statistical power aims to reach a state, where any performance deviations in the benchmarks of models is meaningful. "*About one-third of datasets contain 100k+ questions, which makes them suitable for training and/or fine tuning a deep learning model.*" (Dzendzik et al., 2021). Following the observation in this quote, one could set a minimum criteria on 100k questions in a training dataset. However, it is not the goal of every training dataset to fine tune or specify a model to such a degree. Therefore, this would be an unfair margin to fail smaller datasets on. Instead, a more detailed analysis of both the initial paper that introduces the dataset and the dataset itself is needed to see if the dataset achieves what the paper expects of it. For benchmarking datasets, Bowman and Dahl (2021) argue that it is important to look at what the dataset expects its minimum detectable effect to be, whether it is a change of a 1% absolute accuracy or if it is expected to detect a change of 0.1% absolute accuracy. Bowman and Dahl (2021) argue that it takes a few thousand questions to have a minimum detectable effect, but this can vary wildly depending on the difficulty of the questions.

Points	Description
2.0	**They** have made a detailed analysis to find biases in their data and remove them all.
1.5	**They** have made a detailed analysis to find biases in their data and remove them at best ability.
1.0	**They** have thought about biases and remove them at best ability.
0.5	**Their** trimming manages to remove some biases.
0.0	**Nothing** is done to remove biases.

Table 1: Point distribution for evaluation of attention to biases (max. 2 points).

Points	Description
1.0	**They** have set their dataset up in a framework such that when ever a model is trained or benchmarked on their data, they are measured on relevant benchmarks for biases.
0.5	**They** have a section discussing biases and encourages users to run some matrixes for biases and encourages them to publish their results on these.
0.0	**Nothing** is done to make the community care about biases.

Table 2: Point distribution for evaluation of effort to make community care about biases (max. 1 point).

As for a metric for datasets, it is out of the scope of this paper to make a detailed analysis on each dataset to see what exactly number of questions best fit it. Therefore, we will use a soft lower bound of 100k questions for benchmarking datasets and training datasets alike.

2.3 Grammar checking

Bad grammar is shown to have a negative impact on performance. Belinkov and Bisk (2017) and Si et al. (2020) show that there is a correlation between a high performance and good grammar. In (Belinkov and Bisk, 2017), it is shown that performance in Neural Machine Translation drops significantly when spelling errors are introduced in the data by swapping pairs of characters in words. Si et al. (2020) adapt the same method for MRC models and finds that adjacent character swapping of 7.1% of all words introduces a significant drop in performance at an average of 20.4%.

To find problems in grammar specifically, Dzendzik et al. (2021) propose filtering all words that do not occur in an English dictionary. These words will then be manually checked for slang, names, other languages and unusual words. Furthermore, Richardson et al. (2013) saw an overall performance benefits for proof reading and correction, such as removing slang and punctuation errors. We propose the following metrics to improve a dataset's overall grammar: count the number of non-existing words and slang. Then, we can use proof reading tools (Karyuatry, 2018) and hourly paid linguists who are not biased towards finishing fast in order to validate the punctuation.

2.4 Verification

Verification of a dataset is introduced to verify the correctness of links between sentences, situation model and structure of a question, answer, and input text. For example, it is important that a man is consistently referred to as "he". It is also important to avoid ambiguity, which is an important point that a lot of datasets have emphasized. This enforces that the model is taught against a solid question/answer, rather than a vague, ambiguous, or broad question. Verification of a dataset can be done in several ways, where the norm is using crowdworkers to verify the correctness of questions/answers and data from a dataset (Sugawara et al., 2020). This can be prone to errors if the motivation for these crowdworkers is to churn out as many annotations as possible with no regard for the quality (Sugawara et al., 2020; Richardson et al., 2013; Yang et al., 2015). Crowdworkers might also not be qualified enough to ensure the quality of their own work or spot errors in other's work.

We propose three mechanisms that serve to measure the verification of a dataset. The first mechanism is rating the qualification of the workers in linguistics as shown in Table 3, where only having experts is rated the highest, and not having any criteria is rated the lowest. Second mechanism rates how workers are paid for their contributions. If the workers are paid a stable, hourly pay without a stressful environment of deadlines, they are more likely to be motivated to do a good job. The point distribution for this is shown in Table 4, where there are points given for how the workers are being paid. Hourly pay with no quotas is the best option, which gives 2 points, and rushed workers with no pay is worth 0 points. The last mechanism consists of having those workers peer review questions and contexts in pairs. The point distribution for this is shown in Table 5, where the point system values working in pairs the most, and all work being unchecked valued, is valued the least. All of these mechanisms summed give a number from 0 to 5, where 0 means no effort was put into making sure a good dataset was made, and 5 corresponds to a huge amount of effort being put into making sure a good dataset was built.

Points	Description
1.0	**Only** experts in the field is used.
0.5	**Only** people with a higher education is used.
0.0	**No** restrictions on the workers education is placed.

Table 3: Point distribution for evaluation of worker qualification (max. 1 point).

Points	Description
2.0	**Workers** are not given any timeline or any quota to fulfill, they get paid a hourly pay.
1.5	**Workers** are given hourly pay, but are given instructions to try to achieve some quota.
1.0	**Workers** are paid per work completed.
0.5	**Using** volunteers to work at their own passe.
0.0	**Workers** are not paid and are rushed.

Table 4: Point distribution for evaluation of worker payment (max. 2 point).

2.5 Robustness

Robustness implies that if a model scores high on a benchmark, then the model has a robust in-domain performance. This is achieved by extending the concept of generalizability to all dimensions that are relevant for a given dataset, such as length, difficulty, and in-domain knowledge. Robustness was added to battle the problem of models being over-fitted for a specific dataset, which might be lacking some element of the domain. Some state-of-the-art models achieve close to maximum achievable scores on existing datasets, often outscoring human performance (Bowman and Dahl, 2021). However, that does not necessarily mean a model has a human level of competence. This is emphasized by models getting easily confused in a zero-shot environment (Dunietz et al., 2020). This could be explained by the fact that recent MRC datasets are created to make more difficult domain-specific questions with the premise that if a model can answer more difficult questions, it is ultimately better at the task. The idea of simply more difficult dataset improving models is fundamentally flawed. Some datasets, such as Quoref, are constructed through adversarial filtering. Adversarial filtering filters out questions from the dataset that the model gets right, thus increasing the difficulty of the dataset. However, the idea of adversarial filtering is flawed, as it does not necessarily lead to a better dataset. Bowman and Dahl (2021) argue that adversarial filtering even has a tendency to remove coverage of useful data, simply because the model has already mastered this domain or skill.

Robustness is a difficult criterion to measure, and even humans would have a hard time being able to tell when this is fulfilled. This is because the main dimension is having a broad coverage of knowledge in the expected

Points	Description
2.0	**Having** workers sit together and contribute on the work together.
1.5	**All** questions and answers should be approved by peers.
1.0	**Some** questions and answers are approved.
0.0	**All** questions and answers are taken at face value.

Table 5: Point distribution for evaluation of work review (max. 2 point).

domain, which is not an easy metric to measure. The dataset should also broadly cover the language skills that a model is expected to work with. This might mean including as many skills as possible, as it would be difficult to ensure that models will only be met with benchmarks and tasks inside of a specialized skill. For the purpose of this report, we will calculate a robustness score that will only include the complexity dimension. We will approximate the complexity by using the length, because longer questions and contexts are intuitively more difficult. We take the length of the given context and the given questions. From there, we can check the broadness of complexity by calculating the variance found in sentences and words in questions and contexts. Hence, we define robustness as follows:

$$Robustness = \\ VAR(wl_q) + VAR(wc_q) + VAR(wl_c) + \frac{VAR(wc_c)}{MEAN(wc_c)} ,$$

where wl_q is the length of words in questions, wc_q is the amount of words in questions, wl_c is the length of words in contexts, and wc_c is the amount of words in contexts. The variance of the amount of words in context has to be penalized with the mean in order to not dominate the calculation. More variance in the lengths will lead to a higher score.

Other dimensions, such as language realism, can be measured by checking frequent word usage and comparing this to the frequent word usage in real world English, which can be provided by analysis performed on the Oxford English Corpus. A seemingly impossible task is to check the broad coverage of knowledge in a given domain, and we know of no current way to efficiently test this. We will be discussing potential approaches to fix the problem, but it is difficult to reason about what effects these approaches might have.

3 Dataset analysis

3.1 Selecting datasets

In this section, we will pick a representative quantity of popular datasets for the purpose of analysis based on the criteria. We have picked 31 datasets based mostly on popular discussion in other papers and popularity rankings on `paperswithcode.com` under the tag "Reading Comprehension". These popularity rankings are based on amount of papers that cite the dataset. It is clear from this ranking that SQuAD is extremely overrepresented, with almost four times as many citations as the second highest dataset, MS MARCO. It also has one more benchmark submitted on the website compared to SuperGLUE, which is the second highest in. Many other datasets, including these three, also have their own individual benchmarks, usually located on their own website. We excluded any dataset that was not in English, had incomplete data, or was not available for download to the public.

3.2 Analysis and Results

Disincentivizing bias is done as a detailed analysis of each paper to try to find potential biases they have missed or might have adopted, from crowdworkers or subdatasets which they incorporate. However, as this type of analysis had been done on the first couple of datasets, it became clear that community standards for biases was a lot lower than first expected. Therefore, work started to focus on whether or not the papers actually acknowledged that there could exists potential biases, as seen in Table 6, but this might still have been a bit too optimistic. All analyses for disincentivizing biases and verification can be found in Appendix A.

There are in fact only four datasets which receive above 0 points in disincentivizing biases. Those are AX-g (3.0), AX-b (1.5), ProtoQA (1.25) and SuperGLUE (1.0). AX-g and AX-b are both datasets used to benchmark biases in datasets, it is therefore unsurprising that these received the highest score. SuperGLUE is a composite dataset, so it is good they take the time to see if the datasets they built on are biased in any manner, unlike datasets such as SQuAD, RTE and MRQA, which blindly uses older datasets. SuperGLUE's points come primarily from the fact that they have set a framework up, such that whenever a model is benchmarked on their dataset, they are also measured how biased they are. This is a great initiative and is exactly what we call for the whole industry to do. ProtoQA's points comes from the fact that they have a detailed analysis of which biases exist in their dataset and then they try to remove them to the best of their ability. In addition, they encourage model builders to have biases in mind. Both SuperGlue and ProtoQA are filled with more biases than they are removing and testing for, as they both admit, however what they have done is hopefully the beginning of a more aware industry.

Adequate statistical power is said to be of at least 100k questions on simple datasets. This can be smaller for more advanced datasets, but they still adequate in size to give some meaningful insights into model scores. The average number of questions across all 31 datasets is 388.90k, however if we exclude SuperGLUE and MRQA, as they are composite datasets and therefore counted twice, then the average size is 63.1k. That means the average dataset in general is not big enough for MRC. Datasets such as DuoRC, SQuAD 2.0, NewsQA, TriviaQA, HotpotQA, and SearchQA are adequate of size. At this size, additional questions suffer from dimishing returns, meaning that any significant improvement would take many more questions than it would for smaller datasets. RACE and QuAC are also acceptable. They are not quite at 100k questions, but that is not a definitive line of separation.

The remaining datasets do not come close to 100k questions, they are therefore not adequate of size. However, some of them are still acceptable, as they fall under advanced datasets, such as ComplexWebQuestions with a size of 55k, and NarrativeQA with a size of 49k. A grey area forms around the datasets between 25k - 50k. Here, composition of the dataset and the skills included decide whether or not the given dataset is of adequate size, where one could argue that datasets such as Quoref and WikiQA, with diverse wiki pages as context, would be adequate for training skills such as coreferencing and multi-hop reasoning. All the datasets with less than 10k questions are by no means adequate, such as MCTest, CB, and COPA.

Grammar checking in our evaluation was done by counting the amount of words which do not exist in the dictionary as a percentage of the total amount of words. This lead to most datasets having a percentage lower than 1%, as shown in Table 6. Notable datasets are MCTest and SQuAD 2.0. Those datasaets have an error of 0.08% and 0.13%, which is to be expected, as they are built with having as few grammar errors as possible in mind. On the other hand, there are datasets which have a higher error than 1%, such as RTE and BoolQ. ComplexWebQuestions and WikiQA are the worst, with an error of 3.74% and 3.49% respectively. It should be noted these are QA datasets, which do not have a context. Some of the words in questions or answers might not exist in the dictionary if the words are names for people or places. This is also reflected in datasets which should not have spelling mistakes at all.

Belinkov and Bisk (2017) and Si et al. (2020) state that even small errors lead to substantial performance degradation on models which train on the datasets. As stated in Section 2.3, a 7.1% error degraded the performance of models by an average of 20.4%. Relating this to the results from Table 6 where the highest error is 3.74%, and the lowest is 0.08%, it is clear that the datasets with the error closer to 3.74% significantly degrade the performance of the models which makes those dataset poor in this criteria.

Furthermore, our grammar checking method lacks validation of punctuation and sentence structure. This can be done with a tool such as Grammarly, or another tool that provides similar functionality.

Verification analysis of the papers for all the datasets has shown that a great amount of state-of-the-art datasets are hiring crowdworkers, not only to create their questions and answers, but also to check that all the questions and answers are not ambiguous and can be answered. This also results in that the most received score is 2.5, with 13 datasets receiving this score. What has also become apparent over the analysis of verification is the missing criteria for a proper education for their crowdworkers. This is often good for datasets, since they achieve to get a widespread data which represents the whole population, but misspellings might have been lower if such a restriction had been set.

Since a significant amount of datasets are actually using automatic verification with algorithms, these have not been given the points which they might deserve, as that is not part of the verification metric. However, this is a good initiative and should also be rewarded.

Robustness seems to be reasonably in line with assumptions about which datasets should score highly on this metric. Generally, robustness seems to be somewhat predictable as a combination of the size of the dataset and popularity. Even though MCTest is relatively small, it is of high quality and is very popular, whereas DREAM is five times bigger but has the same robustness. Most datasets seem to be in the range of 10-25, but there are some extremely robust datasets, such as HotpotQA, TriviaQA, and ReCoRD. It would be apparent that composite datasets have a generally high robustness, if not for MRQA. MRQA includes NaturalQuestions in its construction, and NaturalQuestions is by far the largest dataset, with an extremely low robustness score. If MRQA were to exclude NaturalQuestions, and perhaps NewsQA, the robustness would be much higher. SuperGLUE shows that composite datasets seem to be an effective way to increase the robustness.

It is important to note that this method of measurement for robustness is a very simple one. It leaves out a lot of nuance in what it means to be a robust dataset, even in the dimension it tries to measure - complexity. Complex questions and contexts can also be reflected by how difficult and unusual the word usage is in a dataset. There are also specific sentence structures that are harder for NLP models to correctly understand, which is in no way captured by robustness currently. One direct way to inflate robustness is to simply provide an entire Wikipedia page as context versus only relevant information, but that would lead to debate about what is relevant information when checking contexts. Robustness still serves as an indication for variety in questions and contexts, but it is important to highlight its limitations.

Dataset	Year	Q/A source	Data source	B	G	V	S	R
MCTest	2018	Crowdworkers	Crowdwork	0.0	0.08%	2.5	2k	14.51
Quoref	2019	Crowdworkers	Wikipedia	0.0	0.25%	0.75	22k	48.5
RACE	2017	Crowdworkers	English & Chinese exams	0.0	0.41%	2.5	97k	23.29
QuAC	2018	Crowdworkers	Wikipedia	0.0	0.32%	2.5	91k	17.14
NarrativeQA	2018	Crowdworkers	Stories (books, movie scripts)	0.0	0.13%	2.5	49k	21.25
MultiRC	2018	Crowdworkers	Science, news, travel guides, stories	0.0	0.28%	1.5	6k	37.05
CosmosQA	2019	Crowdworkers	Spinn3r Blog	0.0	0.99%	2.5	31k	19.25
QASC	2019	Crowdworkers	Corpus of text	0.0	1.23%	2.5	10k	24.03
DREAM	2019	Crowdworkers	English language exams	0.0	0.34%	2.5	10k	14.9
DuoRC	2018	Crowdworkers	Movie plots (Wikipedia and IMDB)	0.	0.15%	2.0	188k	19.57
MRQA (training set)	2019	Composite dataset	Various data sources.	0.0	0.3%	0.5	1064k	51.81
○ NaturalQuestions (training set)	2019	Crowdworkers	Google search queries	0.0	0.4%	2.5	307k	17.16
○ SQuAD 2.0	2018	Crowdworkers	Wikipedia	0.0	0.13%	2.5	142k	27.64
○ NewsQA	2017	Crowdworkers	CNN	0.0	0.29%	2.5	120k	21.09
○ TriviaQA	2017	Crowdworkers	Crowdsourced trivia	0.0	0.23%	0.5	173k	61.38
○ HotpotQA	2018	Crowdworkers	Wikipedia	0.0	0.37%	1.0	105k	100.82
○ SearchQA	2017	Jeopardy	J! Archive	0.0	0.27%	0.5	216k	46.50
SuperGLUE	2019	Composite dataset	Various data sources.	1.0	0.9%	1.5	162k	120.75
⟁ ReCoRD	2018	Crowdworkers	News articles	0.0	0.91%	3.0	121k	93.38
⟁ AX-b	2019	Authors	FraCaS suite	1.5	1.67%	5.0	0	8.79
⟁ AX-g	2018	Crowdworkers	Manual work inspired by Winograd	3.0	0.15%	5.0	356	20.55
⟁ CB	2019	Crowdworkers	News, fiction, Switchboard	0.0	0.79%	2.5	556	26.25
⟁ COPA	2011	Crowdworkers	Congress Library	0.0	0.21%	2.8	2k	6.76
⟁ RTE	2006	Crowdworkers	News, Wikipedia	0.0	1.68%	1.0	6k	36.23
⟁ WiC	2019	Crowdworkers	WordNet, VerbNet, Wiktionary	0.0	0.42%	2.0	7k	11.45
⟁ WSC	2011	Undisclosed	Fiction books	0.0	0.81%	0.0	2k	9.76
⟁ BoolQ	2019	Crowdworkers	Google search queries	0.0	1.76%	2.0	16k	14.73
♣ ProtoQA	2020	Crowdworkers	Family-feud transcripts	1.25	1.36%	3.25	10k	16.45
♣ CommonsenseQA	2019	Crowdworkers	ConceptNet	0.0	1.19%	2.5	12k	33.14
♣ ComplexWebQuestions	2018	SPARQL	WebQuestionsSP	1.0	3.74%	2.5	55k	15.78
♣ WikiQA	2018	Generation from Wiki	Wikipedia	0.0	3.49%	0.5	29k	11.38

Table 6: Datasets and their criteria scores. B is bias (disincentivizing), G is grammar (checking), V is verification, S is size (statistical power), and R is robustness. Datasets contained in SuperGLUE and MRQA are marked with ⟁ and ○ respectively. Datasets marked with ♣ are datasets without context.

4 Building a dataset

This section describes how we would ideally construct a dataset, which would attempt to fulfill our criteria.

4.1 Getting data

Before data gathering or generation, one needs to take into account what kind of dataset is going to be build, as there are several types to go with. Some examples could be stories, articles, or composite. Once the topic has been decided, one has to think the criteria into the dataset, such as what is the source going to be, does the source comply with the bias, verification and grammar criteria. If not, then what work needs to be done to optimize the data before building it into a dataset.

The very first step towards building the dataset, is to decide on a format for it. Once a format has been decided, a source needs to be picked according to the needs of the dataset. From there, one could get a few passages for context, then make some questions appropriate for the topic and compare it against bias, verification, grammar and robustness to see how much work needs to be done for the given source and if the score for the small test set is as expected before spending too many resources on it. If the source is going to be crowdworkers, then an important factor is considering their level of expertise and cross verifying their work.

To develop a good dataset in the context of disincentivizing biases, it is important to have a thorough analysis of the data, in which all potential biases are discussed, and argued how these could either be fixed or removed. Then, in the best case, all of these would be removed. However, this is often not possible, therefore all biased data that can be removed should be, and the remaining should be mentioned for model developers to be aware of. In the case of a benchmarking dataset, a number of bias benchmarking datasets should be set up so these would be evaluated at the same time as the score on the actual dataset.

4.2 Annotating data

One would utilize crowdworkers for annotating the data, as seen in most other datasets. Experts would be given an hourly pay and would need to peer review other questions and answers in pairs. This is the ideal way for most dataset makers to approach Q/A annotation. The peer review process would probably be the most approachable and cost effective way to ensure verification. The grammar would also be improved during this stage, but it would be easy to simply run the grammar checking on ones' own dataset in order to realistically ensure the best grammar possible. Furthermore, it should mentioned that having the peer review process would make the annotation slower.

4.3 Ensuring Criteria

Probably the most difficult criterion to fully ensure would be robustness. That is because it is almost impossible to tell when the broad coverage of a domain is fulfilled. The only way to be almost certain would be to simply include every paper and book available. If one were to make a dataset that can simply be robust according to our current measurements, it would not be too difficult to split the crowdworkers into intervals, where some crowdworkers would make longer questions than others. They would have to be encouraged, not forced, to write questions of a specific length, in order to still maintain the quality.

Ensuring adequate statistical power is simply a matter of time and funding if the approach is to use human annotators. To ensure that adequate statistical power has been met, there would have to be an appropriate number of questions in relation to the difficulties of them and the precision which is wanted in benchmarking datasets, as described in Section 2.2. In case of a simple dataset, then the very least should be 100k. On more advanced datasets, such as fictional story telling with multi-hop questions and hidden meanings, a smaller quantity would be adequate.

4.4 The case of SQuAD 2.0

In this section, we show how we would have proceeded for building SQuAD 2.0 in comparison to how it was done. We have chosen to structure this section by criteria, this is to give an overview, of how to achieve each criteria. The properties of SQuAD 1.1 is a broad diversity of both answer types and different difficulty of questions in terms of reasoning and question types (Rajpurkar et al., 2016).

When creating SQuAD 2.0 according to our own methods, we would first start with the data source, where the context is Wikipedia and the questions are made by crowdworkers. As SQuAD 2.0 is built on top of SQuAD 1.1 (Rajpurkar et al., 2018), we are already given a lot of data from the start. From there, we would create a set of test questions and test our criteria against this small test dataset.

Bias As described in Section 4.1, it is important not to blindly trust one's own source. Therefore, one of the first things to do would be to check the SQuAD 1.1 dataset for bias. Because SQuAD 1.1 uses Wikipedia pages, it is open for all kinds of biases. Normally, what we would recommend for checking different kind of biases would be to detect them relative to which biases are in focus, along with a detailed analysis of which biases might occur based on where the data is from. However, in this case we can run the models trained on SQuAD 1.1 on bias benchmarking datasets. Next, we would try to run the models on the benchmarking datasets for biases, to get an indication for which biases occur and how dominant they are. When this is done, we would develop a program which would remove all triples from SQuAD 1.1 containing our chosen biases to the of our ability. These might include gender, sexual, religious, racial, and cultural biases. Next, we would try to keep our instructions to the crowdworkers as unbiased as possible, and try to make sure they are written in as neutral language as possible. We would try to get as diverse a group of crowdworkers as possible, as to not exclude any cultures unintentionally. Finally, we had set the dataset up in a framework such that not only the results on our dataset are revealed but also the results on appropriate bias bechmarking datasets.

Grammar in SQuAD 2.0 is inherently checked in a robust way by analyzing the Part-of-Speech tagging and Named-Entity Recognition of all words in SQuAD 1.1 (Rajpurkar et al., 2016). From this, we get various information, such as dates, persons, locations, and other entities. They are counted and compared to the number of words in total for each context/question. This is a very good approach, and the only thing we would do different is to check the spelling and punctuation of all words as well, which we found to be 0.13%. If these mistakes were corrected, it could yield a better result for the models and lower the score in the category other entities, so a larger portion of the wordings are categorized properly.

Verification should be optimized over our tables from Section 2.4 and the amount of resources. Since SQuAD 2.0 is meant to be a very big dataset, it is not possible to only use experts. The next best thing would then be to only allow people with a higher education. However, many datasets make use of the fact that that it is possible to choose crowdworkers from English speaking countries. One has to consider the trade-off between the verification benefits making restriction on level of education / where crowdworkers are from and the potential biases which might emerge from this. However, in this specific situation, we would choose to make a restriction to only allow English speaking countries participate. As for pay, we would give them an hourly pay with no quota to fulfill. We would, however, encourage them to use a specific amount of time per task. We might also chose to hire fast workers for more hours, depending on the progress of the work. It would not seem plausible to have workers sit and work together for a dataset of this size, therefore we would be satisfied with having all questions and answers approved by peers.

Adequate statistical power mostly depends on the dataset maker's resources. In SQuAD 1.1's training set, a total of 87,599 questions were made with 422 articles. When SQuAD 2.0 was released, it featured 43,498 new questions in the training set without an answer from 285 newly introduced articles. In the training set, SQuAD 2.0, has 130,319 questions. We argue in section 2.2 that a simple dataset should have 100k or more questions, which SQuAD 2.0 does not fulfill. Additionally, the dataset has questions, which feature simple lexical lookup, syntactic variation, multiple sentence reasoning, and ambiguous questions. We would not have changed any part of this approach while creating SQuAD 2.0.

Robustness is a difficult criterion, as discussed. In SQuAD 1.1, all paragraphs below 500 characters are stripped away (Rajpurkar et al., 2016). This goes against how we measure robustness, which might also explain their low score in Table 6. We would encourage annotators to diversify their approach to making annotations in the categories of length, sentence structure, and language skills needed.

5 Related work

Critique similar to this work already exists (Sugawara et al., 2020; Dunietz et al., 2020; Schlegel et al., 2020; McCoy et al., 2019). These papers differ in their goals and approaches taken when discussing the current problems with datasets. Sugawara et al. (2020) focuses on what reading comprehension means for humans and what implications this has for models, and Dunietz et al. (2020) creates their criteria according to dimensions through which MRC models can prove they understand small stories. Schlegel et al. (2020) and McCoy et al. (2019) are more concerned with empirical analysis that proves the inaccuracies of current benchmarking.

Ethayarajh and Jurafsky (2020) discuss the utility of NLP leaderboards and proposes that metrics such as prediction cost, model size and robustness should be valued in the leaderboards, as a high ranking model might not be practical to use, and hence provides no utility for a user.

Our focus has been to argue what criteria a dataset needs to fulfill in order to be appropriate for MRC. Perhaps the closest related work to this is Bowman and Dahl (2021). Nevertheless, while Bowman and Dahl (2021) focuses entirely on the criteria and how research can deal with these problems, we analyse to what extend our similar critera are present in currently popular datasets. Our focus is also different in that we focus on training datasets in MRC, whereas they focus on benchmarking in NLP as a whole.

6 Conclusion and future work

We have presented five criteria that serve to measure the current state of MRC datasets. We lay out how these measures can be addressed directly, often through the annotation process. We observe that, according to our criteria, some of the state-of-the-art datasets are not entirely adequate for the purpose of MRC. WikiQA is seemingly one of the worst datasets, according to multiple metrics. To our surprise, SQuAD 2.0 was not as successful in our criteria as its popularity would suggest, as bias consideration is completely absent, and ambiguity is highly present. We present our own approach to building a dataset and compare it to SQuAD 2.0. Through this process, we found that SQuAD in general had considerations that dragged them down in our criteria. SuperGLUE turned out to be one of the best datasets, especially because they show to be a pioneer in disincentivizing biases and robustness due to their variety of tasks. We have made suggestions that can be implemented in the process of creating new datasets that directly address our criteria and some of the shortcomings of popular datasets.

We ask for the community to start putting an effort into disincentivizing biases and try to implement some of our suggestions into their work. One of the more pressing issues that could be worked on would be to quantify robustness in a much more precise manner and also measure other dimensions of robustness, such as broad coverage. It would also be of value to create more bias benchmarking datasets, of which models would be able to be scored, such that when the community is ready to go in this direction, a framework is ready for them.

References

J. Antin and A. Shaw. Social desirability bias and self-reports of motivation: A study of Amazon Mechanical Turk in the US and India. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 2925–2934, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450310154. doi: 10.1145/2207676.2208699. URL https://doi.org/10.1145/2207676.2208699.

Y. Belinkov and Y. Bisk. Synthetic and natural noise both break neural machine translation. *CoRR*, abs/1711.02173, 2017. URL http://arxiv.org/abs/1711.02173.

L. Bentivogli, P. Clark, I. Dagan, and D. Giampiccolo. The seventh PASCAL recognizing textual entailment challenge. In *TAC*. Citeseer, 2011.

T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

M. Boratko, X. Li, T. O'Gorman, R. Das, D. Le, and A. McCallum. ProtoQA: A question answering dataset for prototypical common-sense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1122–1136, Online, Nov. 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.emnlp-main.85.

S. R. Bowman and G. E. Dahl. What will it take to fix benchmarking in natural language understanding? *CoRR*, abs/2104.02145, 2021. URL https://arxiv.org/abs/2104.02145.

E. Choi, H. He, M. Iyyer, M. Yatskar, W.-t. Yih, Y. Choi, P. Liang, and L. Zettlemoyer. QuAC : Question answering in context, 2018. URL https://arxiv.org/abs/1808.07036.

K. W. Church. Emerging trends: I did it, i did it, i did it, but. . . *Natural Language Engineering*, 23: 473–480, 2017. URL https://www.cambridge.org/core/services/aop-cambridge-core/content/view/E04A550C6DFF0154C684888B7B9F68EA/S1351324917000067a.pdf/emerging-trends-i-did-it-i-did-it-i-did-it-but.pdf.

C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions, 2019. URL https://arxiv.org/abs/1905.10044.

P. Dasigi, N. F. Liu, A. Marasović, N. A. Smith, and M. Gardner. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning, 2019. URL https://arxiv.org/abs/1908.05803.

M.-C. de Marneffe, M. Simons, and J. Tonhauser. The CommitmentBank: Investigating projection in naturally occurring discourse. *Proceedings of Sinn und Bedeutung*, 23(2):107–124, Jul. 2019. doi: 10.18148/sub/2019.v23i2.601. URL https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/601.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

J. Dunietz, G. Burnham, A. Bharadwaj, O. Rambow, J. Chu-Carroll, and D. A. Ferrucci. To test machine comprehension, start by defining comprehension. *CoRR*, abs/2005.01525, 2020. URL https://arxiv.org/abs/2005.01525.

M. Dunn, L. Sagun, M. Higgins, V. U. Guney, V. Cirik, and K. Cho. SearchQA: A new q&a dataset augmented with context from a search engine, 2017. URL https://arxiv.org/abs/1704.05179.

D. Dzendzik, J. Foster, and C. Vogel. English machine reading comprehension datasets: A survey. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8784–8804, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-main.693.

K. Ethayarajh and D. Jurafsky. Utility is in the eye of the user: A critique of NLP leaderboards. *CoRR*, abs/2009.13888, 2020. URL https://arxiv.org/abs/2009.13888.

H. Frisch. *Europas Kulturhistorie Bind 1*. Europas kulturhistorie. Politikens Forlag, 1964. ISBN 9788726400908.

L. Huang, R. L. Bras, C. Bhagavatula, and Y. Choi. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning, 2019. URL https://arxiv.org/abs/1909.00277.

M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. *CoRR*, abs/1705.03551, 2017. URL http://arxiv.org/abs/1705.03551.

L. Karyuatry. Grammarly as a tool to improve students' writing quality: Free online-proofreader across the boundaries. *JSSH (Jurnal Sains Sosial dan Humaniora)*, 2(1):83–89,

2018. URL http://www.jurnalnasional.ump.ac.id/index.php/JSSH/article/view/2297/1986.

D. Khashabi, S. Chaturvedi, M. Roth, S. Upadhyay, and D. Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1023. URL https://aclanthology.org/N18-1023.

T. Khot, P. Clark, M. Guerquin, P. Jansen, and A. Sabharwal. QASC: A dataset for question answering via sentence composition, 2019. URL https://arxiv.org/abs/1910.11473.

T. Kočiský, J. Schwarz, P. Blunsom, C. Dyer, K. M. Hermann, G. Melis, and E. Grefenstette. The NarrativeQA reading comprehension challenge, 2017. URL https://arxiv.org/abs/1712.07040.

T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7: 452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026.

G. Lai, Q. Xie, H. Liu, Y. Yang, and E. H. Hovy. RACE: large-scale reading comprehension dataset from examinations. *CoRR*, abs/1704.04683, 2017. URL http://arxiv.org/abs/1704.04683.

H. J. Levesque, E. Davis, and L. Morgenstern. The Winograd schema challenge. KR'12, page 552–561. AAAI Press, 2012. ISBN 9781577355601.

P. S. H. Lewis, P. Stenetorp, and S. Riedel. Question and answer test-train overlap in open-domain question answering datasets. *CoRR*, abs/2008.02637, 2020. URL https://arxiv.org/abs/2008.02637.

N. Markl and S. J. McNulty. Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR, 2022. URL https://arxiv.org/abs/2202.12603.

T. McCoy, E. Pavlick, and T. Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://aclanthology.org/P19-1334.

N. Mehrabi, P. Zhou, F. Morstatter, J. Pujara, X. Ren, and A. Galstyan. Lawyers are dishonest? quantifying representational harms in commonsense knowledge resources. *CoRR*, abs/2103.11320, 2021. URL https://arxiv.org/abs/2103.11320.

Nayuki. Computing wikipedia's internal pageranks, 2019. URL https://www.nayuki.io/page/computing-wikipedias-internal-pageranks.

M. T. Pilehvar and J. Camacho-Collados. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations, 2018. URL https://arxiv.org/abs/1808.09121.

P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. SQuAD: 100,000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250, 2016. URL https://arxiv.org/abs/1606.05250.

P. Rajpurkar, R. Jia, and P. Liang. Know what you don't know: Unanswerable questions for SQuAD. *CoRR*, abs/1806.03822, 2018. URL https://arxiv.org/abs/1806.03822.

M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.acl-main.442.

M. Richardson, C. J. Burges, and E. Renshaw. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA, Oct. 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1020.

M. Roemmele, C. A. Bejan, and A. S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pages 90–95, 2011.

R. Rudinger, C. May, and B. Van Durme. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1609. URL https://aclanthology.org/W17-1609.

A. Saha, R. Aralikatte, M. M. Khapra, and K. Sankaranarayanan. DuoRC: Towards complex language understanding with paraphrased reading comprehension, 2018. URL https://arxiv.org/abs/1804.07927.

V. Schlegel, G. Nenadic, and R. Batista-Navarro. Beyond leaderboards: A survey of methods for revealing weaknesses in natural language inference data and models. *CoRR*, abs/2005.14709, 2020. URL https://arxiv.org/abs/2005.14709.

C. Si, Z. Yang, Y. Cui, W. Ma, T. Liu, and S. Wang. Benchmarking robustness of machine reading comprehension models, 2020. URL https://arxiv.org/abs/2004.14004.

Z. Small. 600,000 images removed from ai database after art project exposes racist bias, 2019. URL https://hyperallergic.com/518822.

S. Sugawara, P. Stenetorp, and A. Aizawa. Prerequisites for explainable machine reading comprehension: A position paper. *CoRR*, abs/2004.01912, 2020. URL https://arxiv.org/abs/2004.01912.

K. Sun, D. Yu, J. Chen, D. Yu, Y. Choi, and C. Cardie. DREAM: A challenge dataset and models for dialogue-based reading comprehension, 2019. URL https://arxiv.org/abs/1902.00164.

A. Talmor and J. Berant. The web as a knowledge-base for answering complex questions, 2018. URL https://arxiv.org/abs/1803.06643.

A. Talmor and J. Berant. MultiQA: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4911–4921, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://aclanthology.org/P19-1485.

A. Talmor, J. Herzig, N. Lourie, and J. Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. *CoRR*, abs/1811.00937, 2018. URL https://arxiv.org/abs/1811.00937.

A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman. NewsQA: A machine comprehension dataset, 2016. URL https://arxiv.org/abs/1611.09830.

A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019. URL https://w4ngatang.github.io/static/papers/superglue.pdf.

Y. Yang, W.-t. Yih, and C. Meek. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics. URL https://aclanthology.org/D15-1237.

Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering, 2018. URL https://arxiv.org/abs/1809.09600.

B. Zhang, P. Williams, I. Titov, and R. Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. *CoRR*, abs/2004.11867, 2020. URL https://arxiv.org/abs/2004.11867.

S. Zhang, X. Liu, J. Liu, J. Gao, K. Duh, and B. V. Durme. ReCoRD: Bridging the gap between human and machine commonsense reading comprehension. *CoRR*, abs/1810.12885, 2018. URL http://arxiv.org/abs/1810.12885.

J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Gender bias in coreference resolution: Evaluation and debiasing methods, 2018. URL https://arxiv.org/abs/1804.06876.

Appendices

A Motivation for scores of dataset, in disincentivizing biases and verification.

A.1 ProtoQA

(Boratko et al., 2020) data comes from the American game show Family-Feud, where by they ask 100 Americans for there answers on a question, each answer for the contestants are then giving points depending on how many on the survey has answered the same.

A.1.1 Disincentivizing biases

Attention to biases 1.0 points - They have thought about biases and remove them at best ability.

by their own admission *"we expect that more nuanced issues of stereotypes are common in the data, but are not as easy to measure with an all-or-nothing measure."* (Boratko et al., 2020).

They have paid no attention to cultural stereotypes.

Effort to make the community care about biases 0.25 points - They encourage model builders to think about biases.

these 0.25 points is generous, since in the following quote they are talking about datasets and not models trained on their dataset, but because they encourage for future work on the subject, these points is given. "Studying the bias in such datasets, and natural stereotypical biases which pre-trained language models have been shown to have (Sheng et al., 2019), would be a valuable topic of future work." (Boratko et al., 2020).

Total score : 1.25 points

A.1.2 Verification

professional linguists This is split between *1.0 points - Only experts in the field is used.* and *0.0 points - No restrictions on the workers education is placed.* because they used crowdworkers where no restrictions were to be found on these criteria. But their work was rated by experts, and than again reclustered with low quality answers removed. We will take the average of these 4 steps, so that it is:

$$(3 * 1.0points + 1 * 0.0points)/4 = 0.75points$$

Offer hourly rate This is not specified in the paper, but it is implied: *1.0 points - Workers are paid per work completed.*.

Pair review *1.5 points - All questions and answers should be approved by peers.*

Total score : 3.25 points

A.2 SQuAD 2.0

Builds on SQuAD (Rajpurkar et al., 2016), without any selection on questions / answers (Rajpurkar et al., 2018). SQuAD again uses Wikipedia sites from the Project Nayuki's Wikipedia's internal PageRank, which is a list of the top 10,000 most important Wikipedia sites as calculated with the PageRank algorithm (Nayuki, 2019). SQuAD selects 536 pages at random among the list, and than select all paragraphs from these sites longer than 500 charecters discarding all images, graphs, tables and figures.

A.2.1 Disincentivizing biases

Attention to biases *0.0 points - Nothing is done to remove biases.*

Effort to make the community care about biases *0.00 points - Nothing is done to make the community care about biases*

Total score : 0.0 points

A.2.2 Verification

professional linguists *0.0 - No restrictions on the workers education is placed*

Offer hourly rate *1.5 points - Workers are given hourly pay, but are given instructions to try to achieve some quota.*

Pair review *1.0 points - Some questions and answers are approved.*

Total score : 2.5 points

A.3 Common SenseQA

(Talmor et al., 2018) uses CONCEPTNET proven biased by (Mehrabi et al., 2021).

A.3.1 Disincentivizing biases

Attention to biases *0.0 points - Nothing is done to remove biases.*

Effort to make the community care about biases *0.00 points - Nothing is done to make the community care about biases*

Total score : 0.0 points

A.3.2 Verification

professional linguists *0.0 - No restrictions on the workers education is placed*

Offer hourly rate *1.0 points - Workers are paid per work completed.*

Pair review *1.5 points - All questions and answers should be approved by peers.*

Total score : 2.5 points

A.4 ComplexWebQuestions

(Talmor and Berant, 2018)

A.4.1 Disincentivizing biases

Attention to biases Since nothing is done to remove biases this paper should get the *0.0 points - Nothing is done to remove biases.* score. However, since they have chosen a dataset which looks to be without biases, after a somewhat thorough search both in the dataset and on related papers nothing problematic was found therefore they receive the rating: *1.0 points - They have thought about biases and remove them at best ability.*

Effort to make the community care about biases *0.00 points - Nothing is done to make the community care about biases.*

Total score : 1.0 points

A.4.2 Verification

professional linguists *0.0 - No restrictions on the workers education is placed.*

Offer hourly rate *1.0 points - Workers are paid per work completed.*

Pair review *1.5 points - All questions and answers should be approved by peers.*

Total score : 2.5 points

A.5 SearchQA

(Dunn et al., 2017)

A.5.1 Disincentivizing biases

Attention to biases *0.0 points - Nothing is done to remove biases.*

Effort to make the community care about biases *0.00 points - Nothing is done to make the community care about biases*

Total score : 0.0 points

A.5.2 Verification

professional linguists The answers are taken from hits on Google, so there is no restriction on education, *0.0 - No restrictions on the workers education is placed.*

Offer hourly rate Since they are using Google results, these are considered volunteers, *0.5 points - Using volunteers to work at their own passe.*

Pair review *0.0 points - All questions and answers are taken at face value*

Total score : 0.5 points

A.6 MCtest

A.6.1 Disincentivizing biases

(Richardson et al., 2013) mentions nothing about trying to avert social biases in their dataset. They use Amazon Mechanical Turk (AMT) to collect data and even though there have been a lot of research finding social biases among the crowdworkers (Antin and Shaw, 2012) (Small, 2019). No thoughts is given to prevent this in the paper nor dataset generations, only dealing with grammar and biases towards what context given to the crowdworkers. Therefore MCtest fully fails this criteria.

0.0 points - Nothing is done to remove biases and *0.00 points - Nothing is done to make the community care about biases.*

Total score : 0.0 points

A.6.2 Verification

professional linguists *0.0 - No restrictions on the workers education is placed*

Offer hourly rate *1.0 points - Workers are paid per work completed..*

Pair review *1.5 points - All questions and answers should be approved by peers.*

Total score : 2.5 points

A.7 WikiQA

(Dunn et al., 2017)

A.7.1 Disincentivizing biases

Attention to biases *0.0 points - Nothing is done to remove biases.*

Effort to make the community care about biases *0.00 points - Nothing is done to make the community care about biases*

Total score : 0.0 points

A.7.2 Verification

professional linguists The answers are taken from hits on Google, so there is no restriction on education, *0.0 - No restrictions on the workers education is placed.*

Offer hourly rate Since they are using Google results, these are considered volunteers, *0.5 points - Using volunteers to work at their own passe.*

Pair review *0.0 points - All questions and answers are taken at face value*

Total score : 0.5 points

A.8 Quoref

(Dasigi et al., 2019)

A.8.1 Disincentivizing biases

Attention to biases They receive their data from Wikipedia and uses Turk to find questions, Turk is previously proven bias. Therefore it is to be expected that the dataset contains some biases. *0.0 points - Nothing is done to remove biases.*

Effort to make the community care about biases *0.00 points - Nothing is done to make the community care about biases*

Total score : 0.0 points

A.8.2 Verification

professional linguists The answers are taken from Wikipedia, so there is no restriction on education, *0.0 - No restrictions on the workers education is placed.*

Offer hourly rate Since they are using Wikipedia pages, these are considered volunteers: *0.5 points - Using volunteers to work at their own passe.*, they are using Turk to find the questions, nothing is said about their salary, so community standard is used here: *1.0 points - Workers are paid per work completed.*

We will take the average of these letting them count the same value.

$$(1.0points + 0.5points)/2 = 0.75points$$

Pair review The questions and answers are validated only by a bot being unable to answer the questions, which do not qualify for any points in this section. *0.0 points - All questions and answers are taken at face value*

Total score : 0.75 points

A.9 SuperGLUE

(Wang et al., 2019) is a collection of a lot of different datasets, more specific all datasets with a '*' next to their name in table 6.

A.9.1 Disincentivizing biases

Attention to biases They take the under laying datasets at face value: *0.0 points - Nothing is done to remove biases.*

Effort to make the community care about biases SuperGlue is set up with a gender bias bechmarking dataset, so when ever a model is ran on SuperGlue it is also ran on this benchmarking. It does however only measure the models decisions of using male or female pronouns but it is a great initiative! *1.0 points - They have set their dataset up in a framework such that when ever a model is trained or benchmarked on their data, they are measured on relevant benchmarks for biases.*

Total score : 1.0 points

A.9.2 Verification

professional linguists They receive their information from under laying datasets, they do not mention any criteria of education for these datasets *0.0 - No restrictions on the workers education is placed.*

Offer hourly rate They receive their information from under laying datasets, and therefore do not use any crowdworkers, therefore an estimate of their own salary is used: *1.5 points - Workers are given hourly pay, but are given instructions to try to achieve some quota.*

Pair review No such criteria is mentioned: *0.0 points - All questions and answers are taken at face value*

Total score : 1.5 points

A.10 TriviaQA

(Joshi et al., 2017)

A.10.1 Disincentivizing biases

Attention to biases Biases is not mentioned in the paper: *0.0 points - Nothing is done to remove biases.*

Effort to make the community care about biases *0.00 points - Nothing is done to make the community care about biases*

Total score : 0.0 points

A.10.2 Verification

professional linguists They recive information from Bing, trivia websites and Wikipedia *0.0 - No restrictions on the workers education is placed.*

Offer hourly rate *0.5 points - Using volunteers to work at their own passe.*

Pair review No such criteria is mentioned: *0.0 points - All questions and answers are taken at face value*

Total score : 0.5 points

A.11 HotpotQA

(Yang et al., 2018)

A.11.1 Disincentivizing biases

Attention to biases Biases is not mentioned in the paper: *0.0 points - Nothing is done to remove biases.*

Effort to make the community care about biases *0.00 points - Nothing is done to make the community care about biases*

Total score : 0.0 points

A.11.2 Verification

professional linguists They receive information from Bing, trivia websites and Wikipedia *0.0 - No restrictions on the workers education is placed.*

Offer hourly rate there is mentions of crowdworkers, it is not mentioned what their pay consists of, so community standardes is used: *1.0 points - Workers are paid per work completed.*

Pair review No such criteria is mentioned: *0.0 points - All questions and answers are taken at face value*

Total score : 1.0 points

A.12 RACE

(Lai et al., 2017)

A.12.1 Disincentivizing biases

Attention to biases Biases is not mentioned in the paper: *0.0 points - Nothing is done to remove biases.*

Effort to make the community care about biases *0.00 points - Nothing is done to make the community care about biases*

Total score : 0.0 points

A.12.2 Verification

professional linguists They use English quizzes developed by English teachers, these are classified as experts in this context: *1.0 points - Only experts in the field is used.*

Offer hourly rate Teachers salary is not mentioned, but is assumed to be the following: *1.5 points - Workers are given hourly pay, but are given instructions to try to achieve some quota.*

Pair review No such criteria is mentioned: *0.0 points - All questions and answers are taken at face value*

Total score : 2.5 points

A.13 NewsQA

(Trischler et al., 2016)

A.13.1 Disincentivizing biases

Attention to biases Biases is not mentioned in the paper: *0.0 points - Nothing is done to remove biases.*

Effort to make the community care about biases *0.00 points - Nothing is done to make the community care about biases*

Total score : 0.0 points

A.13.2 Verification

professional linguists *0.0 - No restrictions on the workers education is placed.*

Offer hourly rate It is not mentioned what the crowdworkers pay is, so community standardes is used: *1.0 points - Workers are paid per work completed.*

Pair review All questions and answers go throw at least one validation: *1.5 points - All questions and answers should be approved by peers.*

Total score : 2.5 points

A.14 QuAC

(Choi et al., 2018)

A.14.1 Disincentivizing biases

Attention to biases Biases is not mentioned in the paper: *0.0 points - Nothing is done to remove biases.*

Effort to make the community care about biases *0.00 points - Nothing is done to make the community care about biases*

Total score : 0.0 points

A.14.2 Verification

professional linguists *0.0 points - No restrictions on the workers education is placed.*

Offer hourly rate *1.0 points - Workers are paid per work completed.*

Pair review This is a difficult distinction, since workers are set to pair up, and have a conversation, in this manner all questions is answered and checked, they are working together, but also, no checking is done directly on each individual answer / question. *An average of 1.0 points is given.*

Total score : 2.0 points

A.15 NarrativeQA

(Kočiský et al., 2017)

A.15.1 Disincentivizing biases

Attention to biases Biases is not mentioned in the paper: *0.0 points - Nothing is done to remove biases.*

Effort to make the community care about biases *0.00 points - Nothing is done to make the community care about biases*

Total score : 0.0 points

A.15.2 Verification

professional linguists *0.0 points - No restrictions on the workers education is placed.*

Offer hourly rate *1.0 points - Workers are paid per work completed.*

Pair review *1.5 points - All questions and answers should be approved by peers.* i

Total score : 2.5 points

A.16 MultiRC

(Khashabi et al., 2018)

A.16.1 Disincentivizing biases

Attention to biases Biases is not mentioned in the paper: *0.0 points - Nothing is done to remove biases.*

Effort to make the community care about biases *0.00 points - Nothing is done to make the community care about biases*

Total score : 0.0 points

A.16.2 Verification

professional linguists *0.0 points - No restrictions on the workers education is placed.*

Offer hourly rate , nothing is metioned abut this and community standards is taken: *1.0 points - Workers are paid per work completed.*

Pair review *1.5 points - All questions and answers should be approved by peers.*

Total score : 2.5 points

A.17 CosmosQA

(Huang et al., 2019)

A.17.1 Disincentivizing biases

Attention to biases Biases is not mentioned in the paper: *0.0 points - Nothing is done to remove biases.*

Effort to make the community care about biases *0.00 points - Nothing is done to make the community care about biases*

Total score : 0.0 points

A.17.2 Verification

professional linguists *0.0 points - No restrictions on the workers education is placed.*

Offer hourly rate , nothing is metioned abut this and community standards is taken: *1.0 points - Workers are paid per work completed.*

Pair review *1.5 points - All questions and answers should be approved by peers.*

Total score : 2.5 points

A.18 QASC

(Khot et al., 2019)

A.18.1 Disincentivizing biases

Attention to biases Biases is not mentioned in the paper: *0.0 points - Nothing is done to remove biases.*

Effort to make the community care about biases *0.00 points - Nothing is done to make the community care about biases*

Total score : 0.0 points

A.18.2 Verification

professional linguists *0.0 points - No restrictions on the workers education is placed.*

Offer hourly rate , nothing is metioned abut this and community standards is taken: *1.0 points - Workers are paid per work completed.*

Pair review *1.5 points - All questions and answers should be approved by peers.*

Total score : 2.5 points

A.19 DREAM

(Sun et al., 2019)

A.19.1 Disincentivizing biases

Attention to biases Biases is not mentioned in the paper: *0.0 points - Nothing is done to remove biases.*

Effort to make the community care about biases *0.00 points - Nothing is done to make the community care about biases*

Total score : 0.0 points

A.19.2 Verification

professional linguists They use English quizzes developed by English teachers, these are classified as experts in this context: *1.0 points - Only experts in the field is used.*

Offer hourly rate Teachers salary is not mentioned, but is assumed to be the following: *1.5 points - Workers are given hourly pay, but are given instructions to try to achieve some quota.*

Pair review *0.0 points - All questions and answers are taken at face value*

Total score : 2.5 points

A.20 DuoRC

(Saha et al., 2018)

A.20.1 Disincentivizing biases

Attention to biases Biases is not mentioned in the paper: *0.0 points - Nothing is done to remove biases.*

Effort to make the community care about biases *0.00 points - Nothing is done to make the community care about biases*

Total score : 0.0 points

A.20.2 Verification

professional linguists *0.0 points - No restrictions on the workers education is placed*

Offer hourly rate *1.0 points - Workers are paid per work completed.*

Pair review *1.0 points - Some questions and answers are approved*

Total score : 2.0 points

A.21 ReCoRD

(Zhang et al., 2018)

A.21.1 Disincentivizing biases

Attention to biases Biases is not mentioned in the paper: *0.0 points - Nothing is done to remove biases.*

Effort to make the community care about biases *0.00 points - Nothing is done to make the community care about biases*

Total score : 0.0 points

A.21.2 Verification

professional linguists *0.0 points - No restrictions on the workers education is placed*

Offer hourly rate Their disclosure on how they are paying their crowdworkers is ambulant, as they meansun what they are paying the crowdworkers on average an hour. This is however also done in papers where they are paying them per completed task, therefore we will take the average of the two scores: *2.0 points - Workers are not given any timeline or any quota to fulfill, they get paid a hourly pay.* and *1.0 points - Workers are paid per work completed.* = 1.5 points

Pair review *1.5 points - All questions and answers should be approved by peers.*

Total score : 3.0 points

A.22 AX-b

(Wang et al., 2019)

A.22.1 Disincentivizing biases

Attention to biases *1.0 points - They have thought about biases and remove them at best ability*

Effort to make the community care about biases *0.5 points - They have a section discussing biases and encourages users to run some matrixes for biases and encourages them to publish their results on these.*

Total score : 1.5 points

A.22.2 Verification

professional linguists *1.0 points - Only experts in the field is used*, the researchers have made it them self.

Offer hourly rate *2.0 points - Workers are not given any timeline or any quota to fulfill, they get paid a hourly pay*, estimate of how the researchers is hired.

Pair review *2.0 points - Having workers sit together and contribute on the work together*, estimate on how they have worked together.

Total score : 5.0 points

A.23 AX-g

(Zhao et al., 2018) is a dataset to test for gender specific biases. It is relevant to know that even though this benchmark receive top score in this paper it is not perfect, they are not analysing their own dataset for biases such as racial, cultural, educational, and so on. The decision to still give this dataset top points in attention to biases is because they still have a detailed section discussing these gender biases.

A.23.1 Disincentivizing biases

Attention to biases *2.0 points - They have made a detailed analysis to find biases in their data and remove them all.*

Effort to make the community care about biases *1.0 points - They have set their dataset up in a framework such that when ever a model is trained or benchmarked on their data, they are measured on relevant benchmarks for biases.* This is the relevant bechmark

Total score : 3.0 points

A.23.2 Verification

professional linguists *1.0 points - Only experts in the field is used*, the researchers have made it them self.

Offer hourly rate average of *2.0 points - Workers are not given any timeline or any quota to fulfill, they get paid a hourly pay*, estimate of how the researchers is hired and *1.0, Workers are paid per work completed.* as is the standard for Turk, used in the paper.

Pair review *2.0 points - Having workers sit together and contribute on the work together*, estimate on how they have worked together.

Total score : 5.0 points

A.24 CommitmentBank

(de Marneffe et al., 2019)

A.24.1 Disincentivizing biases

Attention to biases Biases is not mentioned in the paper: *0.0 points - Nothing is done to remove biases.*

Effort to make the community care about biases *0.00 points - Nothing is done to make the community care about biases*

Total score : 0.0 points

A.24.2 Verification

professional linguists *0.0 points - No restrictions on the workers education is placed.*

Offer hourly rate *1.0 points - Workers are paid per work completed*

Pair review *1.5 points - All questions and answers should be approved by peers*

Total score : 2.5 points

A.25 COPA

(Roemmele et al., 2011)

A.25.1 Disincentivizing biases

Attention to biases Biases is not mentioned in the paper: *0.0 points - Nothing is done to remove biases.*

Effort to make the community care about biases *0.00 points - Nothing is done to make the community care about biases*

Total score : 0.0 points

A.25.2 Verification

professional linguists *1.0 - Only experts in the field is used* and *0.0 points - No restrictions on the workers education is placed* are both used, No retirction workers are used twice as much as experts, experts being the dataset creators.

$$(1 * 1.0 + 2 * 0.0)/3 = 0.3$$

Offer hourly rate *1.0 points - Workers are paid per work completed*

Pair review *1.5 points - All questions and answers should be approved by peers*

Total score : 2.8 points

A.26 RTE

(Bentivogli et al., 2011) is RTE-7. The RTE papers is very focused on the models which will be ran on them, so their focus is not as much on the creation of a dataset.

A.26.1 Disincentivizing biases

Attention to biases Biases is not mentioned in the paper: *0.0 points - Nothing is done to remove biases.*

Effort to make the community care about biases *0.00 points - Nothing is done to make the community care about biases*

Total score : 0.0 points

A.26.2 Verification

professional linguists *0.0 points - No restrictions on the workers education is placed*

Offer hourly rate Standard: *1.0 points - Workers are paid per work completed*

Pair review *0.0 points - All questions and answers are taken at facevalue.*

Total score : 1.0 points

A.27 WiC

(Pilehvar and Camacho-Collados, 2018)

A.27.1 Disincentivizing biases

Attention to biases Biases is not mentioned in the paper: *0.0 points - Nothing is done to remove biases.*

Effort to make the community care about biases *0.00 points - Nothing is done to make the community care about biases*

Total score : 0.0 points

A.27.2 Verification

professional linguists *0.0 points - No restrictions on the workers education is placed.*

Offer hourly rate standard is taken since nothing is specified in the report: *1.0 points - Workers are paid per work completed*

Pair review *1.0 points - Some questions and answers are approved*

Total score : 2.0 points

A.28 WSC

(Levesque et al., 2012), comes with no explanation on how they create a dataset, instead they define their interesting theory, sadly this means they will get no points in this paper.

A.28.1 Disincentivizing biases

Attention to biases Biases is not mentioned in the paper: *0.0 points - Nothing is done to remove biases.*

Effort to make the community care about biases *0.00 points - Nothing is done to make the community care about biases*

Total score : 0.0 points

A.28.2 Verification

professional linguists *0.0 points - No restrictions on the workers education is placed.*

Offer hourly rate Nothing is said so 0 points.

Pair review Nothing is said so 0 points.

Total score : 0.0 points

A.29 BoolQ

(Clark et al., 2019)

A.29.1 Disincentivizing biases

Attention to biases Biases is not mentioned in the paper: *0.0 points - Nothing is done to remove biases.*

Effort to make the community care about biases *0.00 points - Nothing is done to make the community care about biases*

Total score : 0.0 points

A.29.2 Verification

professional linguists *0.0 points - No restrictions on the workers education is placed.*

Offer hourly rate They receive their data from other dataset. *1.0 points - Workers are paid per work completed*

Pair review *1.0 points - Some questions and answers are approved*

Total score : 2.0 points

A.30 NaturalQuestions

(Kwiatkowski et al., 2019)

A.30.1 Disincentivizing biases

Attention to biases Biases is not mentioned in the paper: *0.0 points - Nothing is done to remove biases.*

Effort to make the community care about biases *0.00 points - Nothing is done to make the community care about biases*

Total score : 0.0 points

A.30.2 Verification

professional linguists Combination of experts and no restrictions: so *0.5*.

Offer hourly rate They receive their data from other dataset. *1.0 points - Workers are paid per work completed*

Pair review *1.0 points - Some questions and answers are approved*

Total score : 2.5 points