Assessing the Benchmarking Capacity of Machine Reading Comprehension Datasets

Saku Sugawara∗
University of Tokyo
sakus@is.s.u-tokyo.ac.jp

Kentaro Inui
Tohoku University and RIKEN Center for AIP
inui@ecei.tohoku.ac.jp

Pontus Stenetorp
University College London
p.stenetorp@cs.ucl.ac.uk

Akiko Aizawa
National Institute of Informatics
aizawa@nii.ac.jp

Abstract
Existing analysis work in machine reading comprehension (MRC) is largely concerned with evaluating the capabilities of systems. However, the capabilities of datasets are not assessed for benchmarking language understanding precisely. We propose a semi-automated, ablation-based methodology for this challenge: By checking whether questions can be solved even after removing features associated with a skill requisite for language understanding, we evaluate to what degree the questions do not require the skill. Experiments on 10 datasets (e.g., CoQA, SQuAD v2.0, and RACE) with a strong baseline model show that, for example, the relative scores of the baseline model provided with content words only and with shuffled sentence words in the context are on average 89.2% and 78.5% of the original scores, respectively. These results suggest that most of the questions already answered correctly by the model do not necessarily require grammatical and complex reasoning. For precise benchmarking, MRC datasets will need to take extra care in their design to ensure that questions can correctly evaluate the intended skills.

1 Introduction
Machine reading comprehension (MRC) is a testbed for evaluating natural language understanding (NLU), by letting machines answer questions about given texts (Hirschman et al. 1999). Although MRC could be the most suitable task for evaluating NLU (Chen 2018) and the performance of systems is comparable to humans on some existing datasets Devlin et al. (2019), it has been found that the quality of existing datasets might be insufficient for requiring precise understanding (Jia and Liang 2017). Whereas these analyses are useful to investigate the performance of systems, however, it is still necessary to verify the fine-grained capabilities of datasets for benchmarking NLU.

In the design of MRC datasets, it is implicitly assumed that questions test a cognitive process of language understanding (Sutcliffe et al. 2013). As various aspects of such a process, we can use requisite skills for answering questions such as coreference resolution and commonsense reasoning (Sugawara et al. 2017). Considering skills as metrics would be useful for analyzing datasets. However, for most datasets, the skills required to answer existing questions are not identified, or significant human annotation is needed.

In this study, we propose a semi-automated, ablation-based methodology to analyze the capabilities of MRC datasets to benchmark NLU. Our motivation is to investigate to what extent a dataset allows unintended solutions that do not need requisite skills. This leads to the following intuition: if a question is correctly answered (or solvable) even after removing features associated with a given skill, the question does not require the skill. We show an example of our ablation method in Figure 1. Suppose we wish to analyze a dataset’s capacity to evaluate understanding of texts beyond the information of part-of-speech (POS) tags. To this end, we replace context and question words with POS tags and ID numbers. If a model can still correctly answer this modified question, the question does not necessarily require deep understanding of texts but matching word patterns only. Questions of this kind might be insufficient for developing a model that understands texts deeply as they may reduce models to recognizing superficial word overlaps.

Our methodology uses a set of requisite skills and corresponding ablation methods. Inspired by the computational model of reading comprehension (Kintsch 1988), we exemplify 12 skills on two classes: reading and reasoning (Section 3). Then, we present a large-scale analysis over 10 existing datasets using a strong baseline model (Section 4). In Section 5, we perform a complementary inspection of questions with our ablation methods in terms of the solvability of questions and the reconstructability of ablated features. Finally we discuss, in Section 6, two requirements for developing MRC to benchmark NLU: the control of question solvability and the comprehensiveness of requisite skills.

Our contributions are as follows:
• We propose a semi-automated methodology to analyze the benchmarking capacity of MRC datasets in terms of requisite skills for answering questions.
• With an example set of 12 skills and corresponding input-ablation methods, we use our methodology and examine 10 existing datasets with two answering styles.
• Our analysis shows that the relative performance on questions with content words only, shuffled sentence words,
Figure 1: Example of an ablation test that anonymizes context and question words, applied to a question from SQuAD v1.1 (Rajpurkar et al. 2016) with the correct answer underscored. We found that the baseline model can achieve 61.2% F1 on SQuAD v1.1 even after the anonymization.

and shuffled sentence order averaged 89.2%, 78.5%, and 95.4% of the original performance, indicating that the questions might be inadequate for evaluating grammatical and complex reasoning.

These results suggest that most of the questions currently solved in MRC may be insufficient for evaluating various skills. A limitation of our method is that it cannot draw conclusions regarding questions that remain unsolved, and thus we need to assume a reasonable level of performance for existing models on the dataset to be analysed. Given our findings, we posit that MRC datasets should be carefully designed, e.g., by filtering questions using methods such as the intended NLU skills.

3 Related Work

We briefly survey existing interpretation methods and skill-based analyses for NLU tasks.

Interpretation methods. A challenge with the MRC task is that we do not know the extent to which a successful model precisely understands natural language. To analyze a model’s behavior, existing studies mainly proposed modification of the input. For example, Jia and Liang (2017) showed that the performance of existing models on SQuAD (Rajpurkar et al. 2016) significantly degrades when manually verified distracting sentences are added to the given context. In addition, Feng et al. (2018) demonstrated that MRC models do not necessarily change their predictions even when most question tokens are dropped. Likewise, for the natural language inference task, Gururangan et al. (2018) proposed to hide the premise and to evaluate a model using only the hypothesis. These kinds of analyses are helpful for detecting biases that are unintentionally included in datasets. Nonetheless, to assure that a dataset can evaluate various aspects of NLU, more fine-grained detail is needed than what is allowed by inspection using existing methods.

Skills as units of interpretation. In the topic of interpretable machine learning, Doshi-Velez and Kim (2018) defined the concept of cognitive chunks as the basic units of explanation. In the MRC task, we consider that requisite skills to answer questions are appropriate as such units. A skill-based analysis was conducted by Boratko et al. (2018), who proposed classifications of knowledge and reasoning. Prior to this, Sugawara et al. (2017) also defined a set of 13 requisite skills. However, there are two main issues with these approaches: (i) the human annotation does not necessarily reveal unintended biases that machines can make use of, and (ii) it requires costly annotation efforts. Therefore, we posit that a machine-based analysis is needed and that it should be performed in an automated manner.

3 Dataset Diagnosis by Input Ablation

3.1 Formulation

Our methodology uses a set of requisite skills and corresponding ablation methods. By checking the solvability of questions after applying the ablation methods, we can quantify to what degree the questions allow unintended solutions that do not require the requisite skills. Users can define an arbitrary set of skills to suit their purposes.

We develop a method \( \sigma_i \) that ablates features necessary for the corresponding skill \( s_i \) in a set of requisite skills \( S \). For \( (x, y) \in X \times Y \), whenever \( f(x) = y \), if \( f(\sigma_i(x)) = y \), we recognize that \( x \) is solvable without \( s_i \). Here, \( X \) is the input, \( Y \) is the gold labels, \( (x, y) \) is a pair consisting of an input instance and its gold-standard answer, and \( f \) is a model. When the performance gap between the original and the modified dataset is small, we can infer that most of the questions already solved are solvable without \( s_i \). On the other hand, if the gap is large, a sizable proportion of the solved questions may require \( s_i \).

We note that we cannot draw general conclusions for instances given by conditions other than the abovementioned one. Consider the case where \( f(x) = y \) and \( f(\sigma_i(x)) \neq y \), for example. This only means that \( f \) cannot solve \( x \) without the features ablated by \( \sigma_i \). We cannot conclude that \( x \) requires \( s_i \) in every model because there might exist a model that can solve \( x \) without \( s_i \). However, if there is at least one model \( f \) that solves \( x \) without \( s_i \), this may indicate an unintended way to solve \( x \) while ignoring \( s_i \). Therefore our methodology only requires a single baseline model. Users can choose an arbitrary model for their purposes.
Table 1: Example set of requisite skills \( \{ s_i \} \) and corresponding ablation methods \( \{ \sigma_i \} \). \( f \) is a model and \((x, y)\) is a pair consisting of an input instance and its gold-standard answer. We interpret that for \( x \) s.t. \( f(x) = y \), if \( f(\sigma_i(x)) = y \), then \( x \) is solvable without \( s_i \).

| Comprehension skill \( s_i \) | Ablation method \( \sigma_i \) |
|-------------------------------|-------------------------------|
| 1. Recognizing question words excluding interrogatives | Drop all words except interrogatives (\( wh \)-words and \( how \)) in a question. |
| 2. Recognizing content words | Drop content words in the context. |
| 3. Recognizing function words | Drop function words in the context. |
| 4. Recognizing vocabulary | Anonymize context and question words with their part-of-speech tag. |
| 5. Attending the whole context other than similar sentences | Keep the sentences that are the most similar to the question in terms of unigram overlap and drop the other sentences. |
| 6. Recognizing the word order | Randomly shuffle all words in the context. |

| Reading-class | |
|----------------|--------------------------------|
| 7. Grasping sentence-level compositionality | Randomly shuffle the words in all the sentences except the last token. |
| 8. Understanding of discourse relations | Randomly shuffle the order of the sentences in the context. |
| 9. Performing basic arithmetic operations | Replace numerical expressions (CD tag) with random numbers. |
| 10. Explicit logical reasoning | Drop logical terms such as \( not \), \( every \), and \( if \). |
| 11. Resolving pronoun coreferences | Drop personal and possessive pronouns (PRP and PRPS tags). |
| 12. Reasoning about explicit causality | Drop causal terms/clauses such as \( because \) and \( therefore \). |

| Reasoning-class |
|-----------------|
| 5. Attending the whole context other than similar sentences |
| 6. Recognizing the word order |

3.2 Example Set of Requisite Skills

In this section, we exemplify a skill set that consists of 12 skills along with two classes: reading and reasoning (Table 1). In psychology, there is a tradition of theoretical research on human text comprehension. The construction–integration model (Kintsch 1988) is one of the most acknowledged theories. This model assumes that human text comprehension consists of two processes: (i) construction, in which a reader elaborates concepts and propositions in the text and (ii) integration, in which the reader associates the propositions to understand them consistently. We associate this two-step process with our two classes.

Reading skills. This class deals with six skills of observing and recognizing word appearances, which are performed before reasoning. In MRC, it has been shown that some existing questions can be solved by reading a limited number of words in the question and the context (e.g., by simply attending to context tokens that are similar to those of the questions (Sugawara et al. 2018)). Our goal of this class is, therefore, to ensure that the questions require the reading of the whole question and context uniformly.

Reasoning skills. This class comprises six skills of relational reasoning among described entities and events such as pronoun coreference resolution and logical reasoning. Although these skills are essential for sophisticated NLU, it is difficult to precisely determine whether these types of reasoning are genuinely required in answering a question. Therefore, in this class, we define reasoning-related skills that are performed using the explicit information contained in the context (e.g., \( s_9 \) explicit logical reasoning and \( s_{12} \) reasoning about explicit causality).

In the following, we highlight some of the defined skills. Skill \( s_1 \) is inspired by Feng et al. (2018) and Sugawara et al. (2018). Although their studies proposed dropping question tokens based on their model-based importance or the question length, we instead drop tokens other than interrogatives as interpretable features. Our vocabulary anonymization \( (s_4) \) is mainly inspired by Hermann et al. (2015) where they anonymized named entities to make their MRC task independent of prior knowledge. Our shuffle-based methods \( (s_6 \text{ to } s_8) \) are inspired by existing analyses for other tasks (Khandelwal et al. 2018; Nie, Wang, and Bansal 2019; Sankar et al. 2019). Among them, our purpose for \( s_7 \) is to analyze whether a question requires precise reasoning performed over syntactic and grammatical aspects in each sentence. The remaining skills are described in Appendix A.

Although our proposed definitions can be extended, they are sufficient for the purpose of demonstrating and evaluating our approach. In Section 6, we discuss further directions to develop purpose-oriented skill sets.

4 Experiments and Further Analyses

4.1 Experimental Settings

Datasets. We use 10 datasets. For answer extraction datasets in which a reader chooses a text span in a given context, we use (1) CoQA (Reddy, Chen, and Manning 2019), (2) DuoRC (Saha et al. 2018), (3) HotpotQA (distractor) (Yang et al. 2018), (4) SQuAD v1.1 (Rajpurkar et al. 2016), and (5) SQuAD v2.0 (Rajpurkar, Jia, and Liang 2018). For multiple choice datasets in which a reader chooses a correct option from multiple options, we use (6) ARC (Challenge) (Clark et al. 2018), (7) MCTest (Richardson, Burges, and Renshaw 2013), (8) MultiRC (Khashabi et al. 2018), (9) RACE (Lai et al. 2017), and (10) SWAG (Zellers et al. 2018). For the main analysis, we applied our ablation methods to development sets. We included SWAG because its formulation can be viewed as a multiple-choice MRC task and we would like to analyze the reasons for the high performance reported for the baseline model on this dataset (Devlin et al. 2019). For preprocessing the datasets, we use CoreNLP (Manning et al. 2014).
We specify further details in Appendix B.

Models. As the baseline model, we used BERT-large (Devlin et al. 2019). We fine-tuned it on the original training set of each dataset and evaluated it on a modified development set. For \( \sigma_1 \): vocabulary anonymization, we train the model after the anonymization. For ARC, MCTest, and MultiRC, we fine-tuned a model that had already been trained on RACE to see the performance gained by transfer learning (Sun et al. 2019). We report the hyperparameters of our models in Appendix C. Although we trained the baseline model on the original training set, it is assumed that the upper-bound performance can be achieved by a model trained on the modified training set. Therefore, in Section 4.3, we also see the extent to which the performance improves when the model is trained on the modified training set.

Ablation methods. \( \sigma_2 \) and \( \sigma_3 \): we use a set of stopwords from NLTK (Loper and Bird 2002) as function words. All other words are regarded as content words. We do not drop punctuation. When a token is dropped, it is replaced with an [UNK] token to preserve the correct answer span. \( \sigma_4 \): we use the same ID for the same word in a single given context but different IDs for different contexts. For inflectional words, we anonymize them using their lemma. For example, are would be replaced with @verb2 (= is) if it appeared in Figure 1. In addition, to retain the information of the POS tags, we append its POS tag after each inflectional anonymized word (e.g., is is replaced with @verb{ID} {VBZ}). \( \sigma_6 \): because it is necessary to maintain the correct answer span in the answer extraction datasets, we split the context into segments that have the same length as the gold answer span and shuffle them. \( \sigma_7\) as with \( \sigma_6 \), we split each sentence into segments and shuffle them within each sentence. For \( \sigma_4 \) to \( \sigma_8 \), we averaged the scores over five runs with different seeds and report their variances in Appendix D.

4.2 Results of Reading and Reasoning Skills

We report the results for the skills in Table 2. In the following, \% indicates a relative change from the original F1/accuracy unless specified otherwise. In this section, we describe the notable findings for several skills. The observations for all other skills are explained in Appendix F.

\( s_2 \) and \( s_3 \): recognizing content words and function words. On all datasets, the relative changes for \( s_2 \) were greater than those for \( s_3 \). However, it is remarkable that even with function words alone, the model could achieve 53.0% and 17.4% F1 on CoQA and SQuAD v1.1, respectively. On ARC, RACE, and SWAG, the model showed more than 40% accuracy (>25% of random choice). As for content words only, on all answer extraction datasets, the performance was greater than 78.7% that of the original. On all multiple-choice datasets, it was more than 90.2%. These results imply that most of the questions already solved do not necessarily require grammatical and syntactic reasoning, in which function words are used.

\( s_4 \): recognizing vocabulary beyond POS tags. Surprisingly, for SQuAD v1.1, the baseline model achieved 61.2% F1. It only uses 248 tokens as the vocabulary with the anonymization tags and no other actual tokens. For the other answer extraction datasets, the largest drop (73.6% relative) is by HotpotQA; it has longer context documents than the other datasets, which seemingly makes its questions more difficult. To verify the effect of its longer documents, we also evaluated the baseline model on HotpotQA without distracting paragraphs. We found that the model’s performance was 56.4% F1 (the original performance was 76.3% F1 and its relative drop was 26.1%) which is much higher than that on the context with distracting paragraphs (16.8% F1). This indicates that adding longer distracting documents contributes to encouraging machines to understand a given context beyond matching word patterns.

On the other hand, the performance on the multiple choice datasets was significantly worse; if multiple choices do not have sufficient word overlap with the given context, there is no way to infer the correct answer option. Therefore, this result shows that multiple choice datasets might have a capacity for requiring more complex understanding beyond matching patterns between the question and the context than the answer extraction datasets.

\( s_6 \): recognizing the context word order (context words shuffle). We found that for the answer extraction datasets, the relative performance decreased by 55.6% on average. A moderate number of questions are solvable even with the context words shuffled. We also found that, surprisingly, the average decrease was 21.3% for the multiple choice datasets. The drop on MCTest is more prominent than that on the others. We posit that this is because its limited vocabulary makes questions more context dependent. ARC, in contrast, uses factoid texts, and appears less context dependent.

\( s_7 \): grasping sentence-level compositionality (sentence words shuffle). The performance with sentence words shuffled was greater than 60% and 80% those of the original dataset on the answer extraction and multiple-choice datasets, respectively. This result means that most of the solved questions are solvable even with the sentence words shuffled. However, we should not say that all questions must require this skill; a question can require the performance of some complex reasoning (e.g., logical and multi-hop reasoning) and merely need to identify the sentence that gives the correct answer without precisely understanding that sentence. Nevertheless, if the question is not intended to require such reasoning, we should care whether it can be solved with only a (sentence-level) bag of words. In order to ensure that a model can understand the precise meaning of a described event, we may need to include questions to evaluate the grammatical and syntactic understanding into a dataset.

\( s_8 \): discourse relation understanding (sentence order shuffle). The smallest drop, excluding SWAG, which has one context sentence, was \(-1.3\%\), on SQuAD v1.1. Ex-
Table 2: The performances (%) of the baseline model with the ablation tests on the development set. Values in smaller font are changes (%) relative to the original baseline performance, and the rightmost column (“Rel. avg.”) shows their averages.

| Ablation method \ Dataset | CoQA | DuoRC | Hotpot-QA | SQuAD v1.1 | SQuAD v2.0 | ARC | MCTest | Multi-RC | RACE | SWAG | Rel. avg. |
|---------------------------|------|-------|-----------|------------|------------|-----|--------|----------|------|-------|-----------|
| Answering style           |      |       |           |            |            |     |        |          |      |       |           |
|                          |      |       |           |            |            |     |        |          |      |       |           |
| original                  | 77.4 | 58.4  | 63.0      | 91.5       | 81.9       | 52.7 | 87.8   | 78.0     | 68.8 | 85.4  | 0.0       |
| 1. Q interrogatives only  | 20.1 | 7.4   | 14.2      | 57.6       | 15.0       | 17.5 | 42.0   | 57.6     | 57.6 | 57.6  | -66.8     |
| 2. Function words only    | 53.0 | 31.5  | 58.0      | 97.8       | 17.4       | 110  | 16.6   | 50.2     | 32.2 | 32.2  | -48.6     |
| 3. Content words only     | 60.9 | 13.1  | 47.9      | 18.0       | 52.6       | 118  | 11.8   | 73.5     | 32.6 | 32.6  | -10.8     |
| 4. Vocab. anonymization   | 39.0 | 0.0   | 18.6      | 68.2       | 16.8       | 73.6 | 61.2   | 27.0     | 29.4 | 25.7  | -52.6     |
| 5. Most sim. sent. only   | 32.6 | 7.9   | 35.8      | 38.7       | 16.9       | 73.4 | 68.5   | 57.1     | 50.3 | 50.3  | -30.4     |
| 6. Context words shuff.   | 29.8 | 81.5  | 25.4      | 56.6       | 23.6       | 62.9 | 35.9   | 50.7     | 47.2 | 47.2  | -30.4     |
| 7. Sentence words shuff.  | 53.0 | 13.1  | 35.9      | 38.6       | 43.1       | 32.2 | 62.1   | 32.1     | 64.4 | 64.4  | -21.5     |
| 8. Sentence order shuff.  | 72.2 | 6.8   | 56.1      | 4.0        | 53.7       | 15.6 | 90.3   | 3.1      | 80.7 | 80.7  | -4.6      |
| 9. Dummy numerics          | 75.9 | 1.9   | 57.8      | 1.0        | 60.0       | 5.6  | 89.5   | 2.2      | 78.7 | 78.7  | -4.6      |
| 10. Logical words dropped  | 76.7 | 0.9   | 58.0      | 0.7        | 62.1       | 2.3  | 91.0   | 0.5      | 80.6 | 80.6  | -4.6      |
| 11. Pronoun words dropped  | 76.5 | 1.2   | 57.0      | 2.5        | 63.4       | 4.3  | 91.2   | 0.2      | 81.8 | 81.8  | -4.6      |
| 12. Causal words dropped   | 77.3 | 0.1   | 58.3      | 0.3        | 63.3       | 0.5  | 91.2   | 0.3      | 81.8 | 81.8  | -4.6      |

4.3 Further Analyses

To complement the observations in Section 4.2, we performed further experiments as follows.

The whole question and/or context ablation. To correctly interpret the result for $s_1$, we should know the performance on the empty questions. Likewise, for multiple-choice questions, the performance on the empty context should be investigated to reveal biases contained in the answer options. Therefore, we report the baseline results on the whole question and/or context ablation.$^6$

Our results are reported in Table 3. Although the performance on SQuAD v2.0 was relatively high, we found that the model predicted no answer for all of the questions (in this dataset, almost half of the questions are no answer). The other answer extraction datasets showed a relative drop of 80–90%. This result is not surprising since this setting forces the model to choose an answer span arbitrarily. On the multiple-choice datasets, on the other hand, the accuracies were higher than those of random choice (50% for Multi-RC and 25% for the others), which implies that some bias exists in the context and/or the options.

Training and evaluating on the modified context. A question that was raised during the main analysis is what would happen if the model was trained on the modified input. For example, given that the performance with the content words only is high, we would like to know the upper bound performance when the model is forced to ignore content words also during training. Hence we trained the model with the ablations for the following skills: $s_3$ content words only; $s_4$ context word shuffle; and $s_7$ sentence word shuffle.

The results are reported in the bottom rows of Table 3. On almost all datasets, the baseline model trained on the ablation training set ($s'_3$, $s'_4$, and $s'_7$) displayed higher scores than that on the original training set ($s_3$, $s_4$, and $s_7$). On CoQA, for instance, the relative change from the original score was only −8.3% when the model was trained on $s_3$ content words only. Although $s'_3$ and $s'_7$ with RACE were exceptions, their learning did not converge within the specified number of epochs. We observed that for all datasets the relative upper bounds of performance were on average 92.5%, 80.1%, and 91.8% for $s'_3$, $s'_4$, and $s'_7$, respectively. These results support our observations in Section 4.2, that is, the questions allow solutions that do not necessarily require these skills, and thus fall short of testing precise NLU.

Even without tuning on the ablation training set, however, our methods can make an optimistic estimation of questions that are possibly dubious for evaluating intended skills.

Data leakage in BERT for SWAG. BERT’s performance on SWAG is close to the performance by humans (88.0%). However, the questions and corresponding options for SWAG are generated by a language model trained on the BookCorpus (Zhu et al. 2015), on which BERT’s language model is also pretrained. We therefore suspect that there is severe data leakage in BERT’s language model as reported in Zellers et al. (2019). To confirm this issue, we trained a model without the context (i.e., the first given sentence). The accuracy on the development set, which was also without the context, was 74.9% (a relative decrease of 12.2%). This result suggests that we need to pay more attention to the relations of corpora on which a model is trained and evaluated.

---

$^5$See Appendix G for detailed numbers.

$^6$This approach was already investigated by Kaushik and Lipton (2018). However, there is no overlap in datasets between ours and those they analyzed other than SQuAD v1.1.
but leave further analysis for future work.

Table 3: Results of further analyses: the performance (%) after dropping all question (“Q”) and/or context (“C”) words, and that of the baseline model both trained and evaluated on the modified inputs.

Table 4: Comparison of the human solvability and the baseline model’s performance (%) on questions that are sampled from the ablation tests.

Cases A to C are undesirable for evaluating the systems’ skills, it seems to be useful to distinguish them for further improvement of the dataset creation. We therefore perform the annotation of questions with human solvability; we define that a question is solvable if a reasonable rationale for answering the question can be found in the context.

Results. Table 4 shows the human solvability along with the baseline model’s performance on the sampled questions. The model’s performance is taken from the model trained on the original datasets except for the vocabulary anonymization method. For the content words only on both datasets, the human solvability is higher than the baseline performance. Although these gaps are not significant, we might be able to infer that the baseline model relies on content words more than humans (Case B). Given that the high performance of both humans and the baseline model, most of the questions fall into Case A, i.e., they are easy and do not necessarily require complex reasoning involving the understanding of function words.

For the other three methods, the human solvability is lower than the baseline performance. This result indicates that the questions correctly solved only by the baseline model may contain unintended biases (Case C). For example, the gap in the context word shuffle of RACE is significant (30.0% vs. 75.0%). Figure 2 shows a question that is unsolvable for humans but can be solved by the baseline model. We conjecture that while humans cannot detect biases easily, the model can exploit biases contained in the answer options and their relations to the given context.

5 Qualitative Evaluation

In this section, we qualitatively investigate our ablation methods in terms of the human solvability of questions and the reconstructability of ablated features.

We analyze questions of SQuAD v1.1 and RACE which cover both answering styles and are influential in the community. We randomly sampled 20 questions from each dataset that are correctly solved (100% F1 and accuracy) by the baseline model on the original datasets. Our analysis covers four ablation methods (σ in content words only (involving σ11,12), σ2 vocabulary anonymization, σ3 context word shuffle, and σ4 sentence word shuffle) which provided specific insights in Section 4.

5.1 Human Solvability after the Ablation

Motivation. In Section 4, we observed that the baseline model exhibits remarkably high performance on some ablation tests. To interpret this result, we investigate if a question is solvable by humans and the model. Concretely, the question after the ablation can be (A) solvable by both humans and the model, (B) solvable by humans but unsolvable by the model, (C) unsolvable by humans but solvable by the model, or (D) unsolvable by both humans and the model. For Case A, the question is easy and does not require complex language understanding. For Cases B and C, the model may use unintended solutions because (B) it does not use the same solution as humans or (C) it cleverly uses biases that humans cannot recognize. For Case D, the question may require the skill intended by the ablation method. Although
cases: ablated features are (α) unreconstructable and unnecessary, (β) unreconstructable and necessary, (γ) reconstructable and unnecessary, and (δ) reconstructable and necessary. To verify that ablation methods work, we need to confirm that there are few questions of Case δ. The other cases are not critical to our observations in the main experiment. We therefore perform the annotation with the following queries: (i) are ablated features reconstructable? and (ii) are reconstructable features really necessary for answering? When the answers for both queries are yes, a question is in Case δ. In the annotation, we define that features in a question are reconstructable if the features existing around the rationale for answering the question are guessable. We also require that these features are necessary to decide the answer if the correct answer becomes undecidable without them.

Results. For both datasets, the annotation shows that, not surprisingly, almost all features are unreconstructable in the shuffled sentence/context words and the vocabulary anonymization (except for one example in RACE). When these questions are solvable/unsolvable by humans, we can say that features are unnecessary (Case α) / necessary (Case β) for answering the questions. In contrast, the annotators could guess function words for some questions even if these words are dropped (SQuAD: 55.0% and RACE: 15.0%). The annotation of the necessity also shows that, however, reconstructable features (function words in this case) for all the questions are not necessary to answer them (i.e., Case γ). Therefore, we could not find any question in Case δ. We report the annotation results in Appendix H. It is not easy for the annotator to completely ignore the information of reconstructed features. We leave designing a solid, scalable annotation scheme for future work.

In summary, we found that almost all ablated features are unreconstructable. Although for some questions ablated features are reconstructable for the content words only, these words are not necessarily required for answering the questions. Overall, this result supports our observations in Section 4, i.e., questions already solved in existing datasets do not necessarily require complex language understanding.

6 Discussion

In this section, we discuss two requirements for developing the MRC task as an NLU benchmark.

The control of question solvability. Not to allow the model to focus on unintended objectives, we need to ensure that each question is unsolvable without its intended requisite skill. Therefore, when benchmarking, we first need to identify necessary features whose presence determines the question’s solvability. To identify them, we might need to perform ablation testing with humans. Further, we need to evaluate a model in both regular and ablation settings. This is because a model may detect some biases that enable it to solve the question; such biases can actually be false for humans and may be acquired by the model through overfitting to datasets. Nonetheless, there is a case in which, even if we can identify necessary features, the model can have prior, true knowledge (e.g., world knowledge) of the correct answer. In this case, the model can answer the question without the context. To avoid this circumvention, we may need to evaluate the model on fictional texts.

Comprehensiveness of requisite skills. Another aspect of NLU benchmarking is the comprehensiveness of skills. Our proposed approach can be expanded in two further directions: (i) inner-sentence and (ii) multiple-sentence levels. For (i), we can focus on understanding of specific linguistic phenomena. This includes logical and semantic understanding such as in FraCaS (Cooper et al. 1994) and SuperGLUE (Wang et al. 2019). To investigate particular syntactic phenomena, we might be able to use existing analysis methods (Marvin and Linzen 2018). For (ii), our skills can include complex/implicit reasoning, e.g., spatial reasoning (Weston et al. 2015) and lexically dependent causal reasoning (Sap et al. 2019). Although we do not need to include all of these skills in a single dataset, we need to consider the generalization of models across them.

7 Conclusion

Existing analysis work in MRC is largely concerned with evaluating the capabilities of systems. By contrast, in this work, we proposed an analysis methodology for the benchmarking capacity of datasets. Our methodology consists of input-ablation tests, in which each ablation method is associated with a skill requisite for MRC. We exemplified 12 skills and analyzed 10 datasets. The experimental results suggest that for benchmarking sophisticated NLU, datasets should be more carefully designed to ensure that questions correctly evaluate the intended skills. In future work, we will develop a skill-oriented method for crowdsourcing questions.
Acknowledgments
We would like to thank Max Bartolo, Pasquale Minervini, and the anonymous reviewers for their insightful comments.

References
Boratko, M.; Padigela, H.; Mikkilineni, D.; Yuvraj, P.; Das, R.; McCaullum, A.; Chang, M.; Fokoue-Nkoutche, A.; Kapanipathi, P.; Mattei, N.; Musa, R.; Talamadupula, K.; and Witbrock, M. 2018. A systematic classification of knowledge, reasoning, and context within the ARC dataset. In Proc. of MRQA, 60–70.

Chen, D. 2018. Neural Reading Comprehension and Beyond. Ph.D. Dissertation, Stanford University.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. CoRR abs/1803.05457.

Cooper, R.; Crouch, R.; van Eijck, J.; Fox, C.; van Genabith, J.; Jaspers, J.; Kamp, H.; Finkel, M.; Poesio, M.; Pulman, S.; et al. 1994. FraCaS: A framework for computational semantics. Deliverable 8.62–051.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proc. of NAACL-HLT, 4171–4186.

Doshi-Velez, F., and Kim, B. 2018. Considerations for Evaluation and Generalization in Interpretable Machine Learning. Springer International Publishing, 1st edition.

Feng, S.; Wallace, E.; Grissom II, A.; Iyyer, M.; Rodriguez, P.; and Boyd-Graber, J. 2018. Pathologies of neural models make interpretations difficult. In Proc. of EMNLP, 3719–3728.

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S.; and Smith, N. A. 2018. Annotation artifacts in natural language inference data. In Proc. of NAACL-HLT, 107–112.

Hermann, K. M.; Kocisky, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching machines to read and comprehend. In Proc. of NIPS, 1693–1701.

Hirschman, I.; Light, M.; Breck, E.; and Burger, J. D. 1999. Deep read: A reading comprehension system. In Proc. of ACL, 325–332.

Jia, R., and Liang, P. 2017. Adversarial examples for evaluating reading comprehension systems. In Proc. of EMNLP, 2011–2021.

Kaushik, D., and Lipton, Z. C. 2018. How much reading does reading comprehension require? A critical investigation of popular benchmarks. In Proc. of EMNLP, 5010–5015.

Kandelwal, U.; He, H.; Qi, P.; and Jurafsky, D. 2018. Sharp nearby, fuzzy far away: How neural language models use context. In Proc. of ACL, 284–294.

Khashabi, D.; Chaturvedi, S.; Roth, M.; Upadhyay, S.; and Roth, D. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In Proc. of NAACL-HLT, 252–262.

Kintsch, W. 1988. The role of knowledge in discourse comprehension: A construction-integration model. Psychological review 95(2):163.

Lai, G.; Xie, Q.; Liu, H.; Yang, Y.; and Hovy, E. 2017. RACE: Large-scale reading comprehension dataset from examinations. In Proc. of EMNLP, 796–805.

Loper, E., and Bird, S. 2002. NLTK: The natural language toolkit. In Proc. of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1, 63–70.

Manning, C.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S.; and McClosky, D. 2014. The Stanford CoreNLP natural language processing toolkit. In Proc. of ACL (System Demonstrations), 55–60.

Marvin, R., and Linzen, T. 2018. Targeted syntactic evaluation of language models. In Proc. of EMNLP, 1192–1202.

Min, S.; Zhong, V.; Socher, R.; and Xiong, C. 2018. Efficient and robust question answering from minimal context over documents. In Proc. of ACL, 1725–1735.

Nie, Y.; Wang, Y.; and Bansal, M. 2019. Analyzing compositionality-sensitivity of NLI models. In Proc. of AAAI, 6867–6874.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In Proc. of EMNLP, 2383–2392.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know what you don’t know: Unanswerable questions for SQuAD. In Proc. of ACL, 784–789.

Reddy, S.; Chen, D.; and Manning, C. D. 2019. CoQA: A conversational question answering challenge. TACL 7:249–266.

Richardson, M.; Burges, C. J.; and Renshaw, E. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In Proc. of EMNLP, 193–203.

Saha, A.; Aralikatte, R.; Khapra, M. M.; and Sankaranarayanan, K. 2018. DuoRC: Towards complex language understanding with paraphrased reading comprehension. In Proc. of ACL, 1683–1693.

Sankar, C.; Subramanian, S.; Pal, C.; Chandar, S.; and Bengio, Y. 2019. Do neural dialog systems use the conversation history effectively? an empirical study. In Proc. of ACL, 32–37.

Sap, M.; LeBras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019. ATOMIC: An atlas of machine commonsense for if-then reasoning. In Proc. of AAAI, 3027–3035.

Sugawara, S.; Kido, Y.; Yokono, H.; and Aizawa, A. 2017. Evaluation metrics for machine reading comprehension: Prerequisite skills and readability. In Proc. of ACL, 806–817.

Sugawara, S.; Inui, K.; Sekine, S.; and Aizawa, A. 2018. What makes reading comprehension questions easier? In Proc. of EMNLP, 4208–4219.

Sun, K.; Yu, D.; Yu, D.; and Cardie, C. 2019. Improving machine reading comprehension with general reading strategies. In Proc. of NAACL-HLT, 2633–2643.

Sutcliffe, R.; Peñas, A.; Hovy, E.; Forner, P.; Rodrigo, Á.; Forascu, C.; Benajiba, Y.; and Osenna, P. 2013. Overview of QA4MRE main task at CLEF 2013. Working Notes, CLEF.

Wang, A.; Prukachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In Proc. of NeurIPS, 3261–3275.

Weston, J.; Bordes, A.; Chopra, S.; and Mikolov, T. 2015. Towards AI-complete question answering: a set of prerequisite toy tasks. In the International Conference on Learning Representations.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Proc. of EMNLP, 2369–2380.
A Our Defined Requisite Skills

Reading skills. As $s_2$ and $s_5$, we propose limiting the information available in the context by dropping content and function words respectively, which is intended to ascertain the extent to which a question depends on the given word type (e.g., a preposition in before a time-related expression for a when question). Skill $s_5$ provides a heuristic of the relative levels of attention between a question and the context. Skill $s_6$ is used to ensure that a model can extract the information conditioned on the word order.

Reasoning skills. Skill $s_8$ is for the understanding of discourse relations between adjacent sentences, which relies on information given by the sentence order in the context. When we shuffle the sentence order, various relations, such as causality and temporality, are expected to be broken. Skills $s_9$ to $s_{12}$ are defined more specifically; we drop tokens that explicitly emphasize important roles in specific skills such as if and not in logical reasoning.

B Experimental Details

In this section, we provide details of the specifications used in our experiments.

Datasets. For CoQA, since this dataset allows for yes/no/unknown questions, we appended these answers to the end of the context. These special words were not allowed to be dropped. Additionally, we appended the previous question-answer pair prior to the current question so that the model can consider the history of the QA conversation. To compute the performance on SQuAD v2.0, we used the best F1 value that was derived from the predictions with a no-answer threshold of 0.0. For DuoRC, we used the ParaRC dataset (the official preprocessed version provided by the authors). When training a model on DuoRC and HotpotQA, we used the first answer span; i.e., the document spans that have no answer span were not used in training. For MCTest and RACE, we computed accuracy by combining MC160 with MC500 and Middle with High, respectively. For MultiRC, which is allowed to have multiple correct options for a question, we cast a pair consisting of a question and one option as a two-option multiple choice (i.e., whether its option is true or false) and computed the micro-averaged accuracy for the evaluation. The SWAG dataset is a multiple-choice task of predicting which event is most likely to occur next to a given sentence and the subject of the second sentence as the question. To compute F1 scores for the answer extraction datasets, we used the official evaluation scripts provided for the answer extraction datasets.

Ablation methods. For $\sigma_3$ vocabulary anonymization, we used the tags as shown in Table 5 and @other tags for the other POS tags. For $\sigma_{10}$ logical words dropped, as logic-related terms, we used the following: all, any, each, every, few, if, more, most, no, nor, not, other, some, some, and then. For $\sigma_2$ causal words dropped, as causality-related terms, we used the following: as, because, cause, since, therefore, and why. For $\sigma_3$ training with content words only, we dropped function words as well as punctuation marks so that the model would see only content words.

We show examples of questions for the ablation method $\sigma_4$ in Figure 3.

C Hyperparameters of the Baseline Model

Hyperparameters used in the experiments are shown in Table 6. We used $\sigma$, $\nu$, $\tau$, and $\eta$ for content words.

Table 6: Examples of anonymization tags and corresponding POS tags (OntoNotes 5 version of Penn Treebank tag set). We use $@noun$, $@verb$, $@adj$, $@adv$, and $@number$ for content words.

| Dataset      | $d$ | $b$ | $l_r$ | $e_p$ |
|--------------|-----|-----|-------|-------|
| CoQA         | 512 | 24  | $3 \times 10^{-5}$ | 2     |
| DuoRC        | 512 | 24  | $3 \times 10^{-5}$ | 2     |
| HotpotQA     | 512 | 24  | $3 \times 10^{-5}$ | 2     |
| SQuAD v1.1   | 384 | 24  | $3 \times 10^{-5}$ | 2     |
| SQuAD v2.0   | 384 | 24  | $3 \times 10^{-5}$ | 2     |
| ARC          | 384 | 24  | $1 \times 10^{-5}$ | 4     |
| MCTest       | 512 | 16  | $2 \times 10^{-6}$ | 4     |
| MultiRC      | 512 | 24  | $2 \times 10^{-5}$ | 4     |
| RACE         | 512 | 32  | $1 \times 10^{-5}$ | 4     |
| SWAG         | 128 | 32  | $1 \times 10^{-5}$ | 4     |

Table 7: Ablation results with variances in parentheses for shuffle-related skills ($s_6$, $s_7$, and $s_8$) for five different runs.
Table 8: Statistics of the datasets examined and average numbers of tokens dropped by our ablation methods $\sigma_i$ ($i = 1, 2, 3, 5, 9, \ldots, 12$). The tokens are counted after tokenization of the punctuation. Values in smaller font denote the proportion (%) of questions that contain dropped tokens.

D Performance Variances in Shuffle Methods

We report the variance for shuffling methods $s_i$ context words shuffle, $s_7$ sentence words shuffle, and $s_8$ sentence order shuffle in Table 7.

E Statistics of the Examined MRC Datasets

Table 8 shows the statistics for the examined MRC datasets.

F Full Observations of the Main Results

In this appendix, we describe the results for the reading and reasoning skills not mentioned in Section 4.2.

a1: recognizing question words. For the first four answer-extraction datasets, the performance decreased by more than 70%. For the multiple-choice datasets, the performance decreased by an average of 23.9%.

a2: attending to the whole context other than similar sentences. Even with only the most similar sentences, the baseline models achieved a performance level greater than half their original performances in 8 out of 10 datasets. In contrast, HotpotQA showed the largest decrease in performance. This result reflects the fact that this dataset contains questions requiring multi-hop reasoning across multiple sentences.

a3–a5: various types of reasoning. For these skills, we can see that the performance drops were small; given that the drop for $s_3$ recognizing content words alone was under 20%, we can infer that specific types of reasoning might not be critical for answering the questions. Some types of reasoning, however, might play an essential role for some datasets: $s_9$ numerical reasoning in HotpotQA (whose questions sometimes require answers with numbers) and $s_{11}$ pronoun coreference resolution in DuoRC (consisting of movie scripts).

G Detailed Results of SQuAD v2.0

We report the ablation results for has-answer and no-answer cases in SQuAD v2.0 in Table 9.

H The Annotation Results

Table 10 shows the frequency of questions for Cases $\alpha$ to $\delta$ for SQuAD v1.1 and RACE. Ablated features are ($\alpha$) reconstructable and unnecessary, ($\beta$) unreconstructable and necessary, ($\gamma$) reconstructable and unnecessary, and ($\delta$) reconstructable and necessary. Questions for Case $\delta$ are problematic for interpreting our main observations.
Having a great collection of books at home doesn’t really mean that you are a person who has a passion for literature and reading. It can be a family inheritance or it can be just to impress people around you by the fact that you are a person of culture. On the other hand, there are many persons who cannot afford to buy books, because some of them are quite expensive, but who usually go to libraries and spend hours reading something that interests them a lot, or just borrow books to home. From my point of view, literature is very important in our life. For example, reading is a means of gaining culture and enriching our knowledge in different areas. It can help us have a great imagination and it makes things easier when it comes to make compositions on different themes. [...]