Moderate to Answer MRC Examples with High Variability are Effective in MRC Model Training

Hongyu Li, **Tengyang Chen**, **Shuting Bai**, **Tingxuan Li**, **Takehito Utsuro**
Graduate School of Systems and Information Engineering, University of Tsukuba, Japan

1 Introduction

The Machine Reading Comprehension (MRC) task locates the best corresponding natural language answer when provided a question and its related context. In recent years, MRC models using neural networks have been proposed for SQuAD [8, 9], which is a large-scale, high-quality English MRC dataset. Most recent neural network based MRC models have outperformed human performance [1].

Among those existing work, to analyze the difficulty of several popular MRC benchmarks such as bAbI [11], SQuAD [8], CBT [3], CNN [2] and Who-did-What [7], Kaushik and Lipton [4] established sensible baselines for these datasets, and found that question-only and context-only (which is referred to as *passage-only* in Kaushik and Lipton [4]) models often performs surprisingly well. In particular, context-only models achieve over 50% accuracy on 14 out of 20 bAbI tasks, and as for CBT, only the last one of the 20 sentences provided as a context is necessary to achieve a comparable accuracy. They also indicated that SQuAD is designed more carefully than other datasets and achieved F1 scores of only 4% and 14.8% respectively on question-only and context-only models, which are relatively lower. Kaushik and Lipton [4] demonstrated that published MRC datasets should characterize the level of difficulty, and specifically, the extent to which questions and contexts are essential. Moreover, they also claimed that follow-up papers reporting improvements ought to report performance both on the full task and variations omitting questions and contexts.

In view of the point demonstrated in Kaushik and Lipton [4], our prior work [5] concentrates more on the difficulty of every single MRC example. Based on pre-trained BERT, we proposed a method of splitting the 87,600 SQuAD1.1 training examples comprised of 12,500 "easy to answer" ones and 75,100 "hard to answer" ones, and found that when using 12,500 examples of each class to fine-tune BERT on MRC task respectively, the "hard to answer" ones significantly outperform "easy to answer" ones on training effectiveness. However, it was also pointed out that the performance of "hard to answer" examples is comparable with that of examples randomly sampled from the training examples of SQuAD1.1.

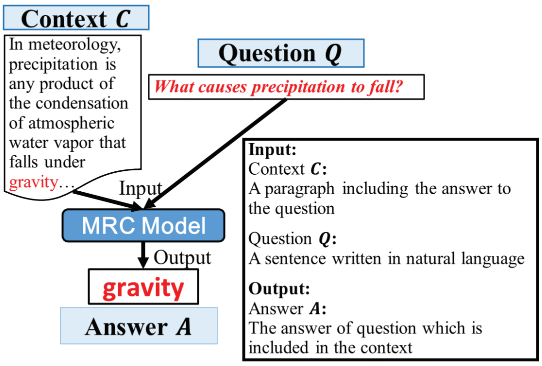

Context C
In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity...
Question Q
What causes precipitation to fall?
Input
MRC Model
Output
gravity
Answer A
Input:
Context C:
A paragraph including the answer to the question
Question Q:
A sentence written in natural language
Output:
Answer A:
The answer of question which is included in the context

Figure 1: An MRC Model using Neural Networks

Swayamdipta et al. [10] proposed *DataMaps*—a general framework of identifying three regions, namely, ambiguous, easy to learn, and hard to learn within a dataset, and applied the framework to several tasks such as natural language inference and sentence-level machine reading comprehension. It is concluded that ambiguous instances are useful for high performance, easy to learn instances aid optimization, and hard to learn instances correspond to data errors.

Following our prior work [5], this paper further proposes a method that splits the MRC examples into three classes of "easy to answer", "moderate to answer", and "hard to answer"[1]. Given the MRC dataset SQuAD1.1 (where each MRC example is denoted as the tuple $\langle Q, C, A\rangle$ of the question Q, the context C, and the answer A) and the pre-trained model RoBERTa [6][2], we apply a 10-fold cross-validation on the 87,600 SQuAD1.1 training examples comprised of 13,400 "easy to answer", 49,300 "moderate to answer", and 25,000 "hard to answer" examples. From the comparison results, the followings are our significant findings. (1) Based on the performance of training the RoBERTa MRC models with 13,400 "easy to an-

[1] The definition of "hard to answer" examples is different from that in our prior work [5].

[2] Our implementation uses the Huggingface Transformers library [12], and all of our models are fine-tuned based on RoBERTa-base with the number of epochs as 2, batch size as 8, learning rate as 0.000015, and the maximum sequence length as 384.

Table 1: Splitting the MRC Examples into "Answerable" and "Unanswerable" Examples with the Corresponding Statistics

		Training set	
		Questions used tr_i	Questions not used ${tr_i}^{Q=\emptyset}$
Test set	questions used s $(\in ts_i)$	Answerable examples $a^1(ts_i)$ Unanswerable examples $ua^1(ts_i)$ $\|a^1\| = 61,909$ $\|ua^1\| = 25,690$	Answerable examples $a^2(ts_i)$ Unanswerable examples $ua^2(ts_i)$ $\|a^2\| = 12,666$ $\|ua^2\| = 74,933$
	questions not used $s^{Q=\emptyset}$ $(s \in ts_i)$	Answerable examples $a^3(ts_i)$ Unanswerable examples $ua^3(ts_i)$ $\|a^3\| = 2,134$ $\|ua^3\| = 85,465$	Answerable examples $a^4(ts_i)$ Unanswerable examples $ua^4(ts_i)$ $\|a^4\| = 9,712$ $\|ua^4\| = 77,877$

swer", "moderate to answer", "hard to answer" examples, as well as those randomly sampled from the complete SQuAD1.1 training examples, the one trained with "moderate to answer" examples outperforms the other three. (2) In comparison with the model trained with 13,400 "moderate to answer" examples, another model trained with the examples with high variability of confidence within "moderate to answer" examples shows further performance improvement.

2 Splitting MRC Examples into "Easy to Answer", "Moderate to answer", and "Hard to Answer" Classes

Following the procedure for detecting MRC examples as answerable without a question demonstrated in the previous section, we similarly split the 87,600 SQuAD1.1 training examples into "easy to answer", "moderate to answer", and "hard to answer" classes. With the 10-fold cross-validation procedure illustrated in Figure 2, we obtain the following four types of evaluation results, where there exist two choices each for both training and evaluation:

(i) The MRC model is trained with the training examples that *include* questions used as they are or *do not include* questions (i.e., "with-question training" or "without-question training").

(ii) The trained MRC model is evaluated against the MRC test examples *with* or *without* questions (i.e., "with-question" evaluation or "without-question" evaluation).

The 10-fold cross-validation results in the three classes of "easy to answer" (approximately 13,400 examples), "moderate to answer" (approximately 49,300 examples), and "hard to answer" (approximately 25,000 examples).

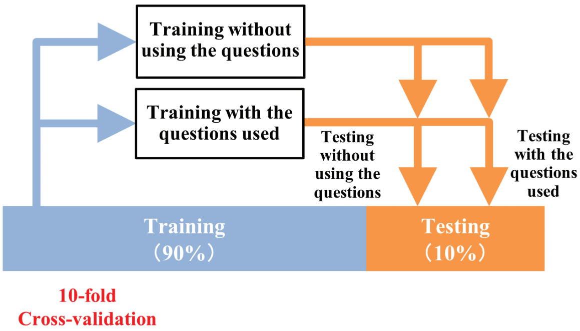

Training without using the questions
Training with the questions used
Testing without using the questions
Testing with the questions used
Training (90%)
Testing (10%)
10-fold Cross-validation

Figure 2: Four Types of Evaluation through 10-fold Cross-Validation for Splitting the MRC Examples into "Easy to Answer", "Moderate to answer", and "Hard to Answer" Classes

Detailed Procedure

The SQuAD1.1 dataset is composed of approximately 100,000 MRC examples that use 23,215 paragraphs extracted from 536 Wikipedia articles as context. With these contexts, questions and answers are annotated through crowd sourcing to generate the complete 100,000 MRC example set. From these examples, we apply N-fold cross-validation (N=10 in this paper) to the set U of the MRC training examples collected from 442 out of the 536 Wikipedia articles.

Before the N-fold cross-validation, we first divide the 442 Wikipedia articles into disjoint N subsets. From the i-th ($i = 1, \ldots, N$) subset of Wikipedia articles, we obtain the i-th test set ts_i of the MRC examples. and the i-th training set of the MRC examples is obtained as the set tr_i of the remaining MRC examples. Then, the set U of the complete SQuAD1.1 training examples is represented as

$$U \;=\; \bigcup_{i=1,\ldots,N} ts_i \quad \left(\, ts_i \cap ts_j = \emptyset \quad (i \neq j) \, \right)$$

As shown in Table 1, from the i-th training set tr_i of the MRC examples, each of which contains a question, another training set ${tr_i}^{Q=\emptyset}$ of the MRC examples is obtained by removing the question Q from each example. So, each MRC example in the obtained training set ${tr_i}^{Q=\emptyset}$ now has an empty question. Similarly, from a test MRC example $s = \langle Q, C, A \rangle$ in the i-th test set ts_i of the MRC examples that contains a question, another test MRC example $s^{Q=\emptyset}$ is obtained by removing its question Q from s. So, the obtained test MRC example $s^{Q=\emptyset}$ has an empty question. Here, we define a Boolean predicate *answerable* which classifies if the given test example s is "answerable" or "unanswerable" by the MRC model $m(tr)$ trained with training set tr, and is defined according to if the predicted answer $\widehat{A}$ is the same as the reference answer A as

$$answerable\Big(m(tr),\ s\Big) \quad = \quad \begin{cases} 1 & (\widehat{A} = A) \\ 0 & (\widehat{A} \neq A) \end{cases}$$

By pairing the two training sets tr_i and ${tr_i}^{Q=\emptyset}$ from the MRC examples and the two test MRC examples s and $s^{Q=\emptyset}$, as shown in Table 1, a resulting four pairs of training sets from the MRC examples and a test MRC example can be examined as to if the given designated test MRC example is "answerable" or "unanswerable" by the MRC model trained with the designated training set.

Finally, in each of these four pairs, the set ts_i of the test MRC examples is split into the set $a^{\alpha}(ts_i)$ of answerable test MRC examples and $ua^{\alpha}(ts_i)$ of unanswerable test MRC examples, according to ($\alpha = 1, 2, 3, 4$)

$$\begin{aligned} & a^1(ts_i) = \\ & \quad \Big\{ s \in ts_i \Big| answerable\Big(m(tr_i), s\Big) = 1 \Big\} \\ & a^2(ts_i) = \\ & \quad \Big\{ s \in ts_i \Big| answerable\Big(m({tr_i}^{Q=\emptyset}), s\Big) = 1 \Big\} \\ & a^3(ts_i) = \\ & \quad \Big\{ s \in ts_i \Big| answerable\Big(m(tr_i), s^{Q=\emptyset}\Big) = 1 \Big\} \\ & a^4(ts_i) = \\ & \Big\{ s \in ts_i \Big| answerable\Big(m({tr_i}^{Q=\emptyset}), s^{Q=\emptyset}\Big) = 1 \Big\} \end{aligned}$$

$$ua^{\alpha}(ts_i) \quad = \quad ts_i - a^{\alpha}(ts_i) \quad (\alpha = 1, 2, 3, 4)$$

The sets $a^{\alpha}(ts_i)$ ($\alpha = 1, 2, 3, 4$) of "answerable" test MRC examples are obtained by evaluating the MRC model trained with the training sets tr_i (with questions) or ${tr_i}^{Q=\emptyset}$ (without questions) against s (with a question) or $s^{Q=\emptyset}$ (without a question). We define the set E of "easy to answer" MRC examples as the union of the three sets $a^{\alpha}(ts_i)$ ($\alpha = 2, 3, 4$) of "answerable" test MRC examples. For these, we collect the "answerable" test MRC examples over the cases with questions removed either from the training or test MRC examples ($a^1(ts_i)$ is excluded because the questions are used in both the training and test MRC examples). The set H of "hard to answer" MRC examples is defined as the intersection of $ua^{\alpha}(ts_i)$ ($\alpha = 1, 2, 3, 4$), which indicates these ones are found as "unanswerable" over all of the four cases. The set M of "moderate to answer" examples is subsequently defined as the complement set of $E \cup H$ as below: [3]

$$\begin{aligned} E &= \bigcup_{i=1,\ldots,N} a^2(ts_i) \vee a^3(ts_i) \vee a^4(ts_i) \\ H &= \bigcup_{i=1,\ldots,N} ua^1(ts_i) \wedge ua^2(ts_i) \wedge ua^3(ts_i) \wedge ua^4(ts_i) \\ M &= U - (E \vee H) \end{aligned}$$

Consequently, the set U of the complete SQuAD1.1 training examples is split into the set E of 13,357 "easy to answer" examples, the set M of 49,251 "moderate to answer" ones, and the set H of 24,991 "hard to answer" ones.

Table 2: Number of Examples in Each Training Set

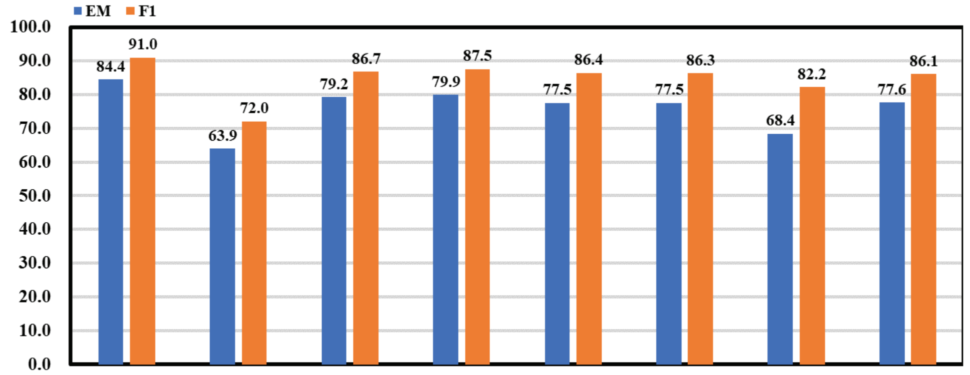
Training set	Number of examples
U: training set of SQuAD1.1	87,599
E: "easy to answer" examples	13,357
M: "moderate to answer" examples	49,251
H: "hard" examples	24,991
M_{sml}: examples randomly sampled from M	13,357
M_{var}: examples with the highest variability within M	
A_{sml}: examples randomly sampled from A	
A_{var}: examples with the highest variability within A	
H_{sml}: examples randomly sampled from H	
U_{sml}: examples randomly sampled from U	

3 Effectiveness of "Moderate to Answer" Examples in MRC Model Training

We next evaluate the effectiveness of "easy to answer", "moderate to answer" and "hard to answer" MRC examples based on the performance of each class when used for the MRC model training. As a baseline, we apply the framework of Swayamdipta et al. [10] on SQuAD1.1, resulting in "ambiguous", "easy to learn", and "hard to learn" examples selected from the training examples of SQuAD1.1. Specifically, we fine-tune RoBERTa model on MRC task for 2 epochs[4] with the training examples of SQuAD1.1, and for each example, we yield the confidence and the variability of the confidence across the 2 epochs. Then, the 29,075 (33% of U) training examples with the highest variability are selected as the set A ("ambiguous"), which outperforms the 29,075 training examples with the highest average confidence ("easy to learn") and the 29,075 training

[3] Over the set U of the complete SQuAD1.1 training examples, the set a^{α} of "answerable" examples and the set ua^{α} of "unanswerable" examples are defined as $a^{\alpha} = \bigcup_{i=1,\ldots,N} a^{\alpha}(ts_i)$, $ua^{\alpha} = U - a^{\alpha}$ ($\alpha = 1, 2, 3, 4$), where the number of examples in each set is provided in Table 1.

[4] We have compared the results among 2∼5 epochs of fine-tuning, and the "ambiguous" examples of 2 epochs of fine-tuning perform the best.



Training	U (87,599)	E (13,357)	M_{sml} (13,357)	M_{var} (13,357)	A_{sml} (13,357)	A_{var} (13,357)	H_{sml} (13,357)	U_{sml} (13,357)
	Training data of SQuAD1.1 U= E∪H	“Easy to answer” examples	Examples randomly sampled from M	Examples with highest variability within M	Examples randomly sampled from A	Examples with highest variability within A	Examples randomly sampled from H	Examples randomly sampled from U

Figure 3: Evaluation Results on the Development Set of SQuAD1.1 where EM is an Exact Match, and F1 is the Macro-Average of the F1 Score per Example

examples with the lowest average confidence (“hard to learn”).

The sets shown in Table 2 are evaluated as the MRC model training examples. In addition to the set E, we evaluate the sets M_{sml}, H_{sml} and U_{sml} of $|E| = 13,557$ MRC examples randomly sampled from M, H, and U, respectively. These sets are intended to directly compare the effectiveness of the “easy to answer”, “moderate to answer”, “hard to answer”, and (randomly sampled) SQuAD1.1 training examples by restricting the numbers of the training examples to be the same. To compare our results with the baseline, we also evaluate the set A_{sml} of $|E| = 13,557$ MRC examples randomly sampled from A. Moreover, we evaluate M_{var} and A_{var} which are the sets of $|E| = 13,557$ MRC examples with the highest variability within M and A, respectively. All these sets are used to fine-tune the RoBERTa pre-trained model on the MRC task, and the development set of SQuAD1.1 is used as the test set for each evaluation. For the evaluation measures, we utilize the exact match (EM), which is defined as the rate of examples with a predicted answer that exactly matches the reference answer. The macro average of the F1 score is calculated from the precision and recall between the token sequences of the predicted and reference answers.

Figure 3 compares the performance of these sets of MRC training examples. Within the three sets of “easy to answer”, “moderate to answer” and “hard to answer” examples, M_{sml} significantly outperforms E and H_{sml}. Compared with the performance of the baseline’s “ambiguous” examples A_{sml}, which is comparable with that of U_{sml}, our “moderate to answer” examples M_{sml} outperforms the set U_{sml} with a statistically significant ($p<0.01$) difference. Furthermore, the set M_{var} of examples with the highest variability of confidence within M outperforms the set M_{sml} with a statistically significant (p<0.1) difference, which also outperforms all other sets of $|E| = 13,557$ examples with a statistically significant (p<0.01) difference. This suggests that “moderate to answer” examples with high variability are effective in MRC Model training.

4 Conclusion

We proposed a method based on RoBERTa [6] that splits the training examples from the MRC dataset SQuAD1.1 into classes of “easy to answer”, “moderate to answer”, and “hard to answer”. Experimental evaluation results of comparing the four models, which are respectively trained only with the “easy to answer”, “moderate to answer”, “hard to answer” examples and examples randomly sampled from the complete SQuAD1.1 training examples, demonstrate that the one trained with “moderate to answer” examples outperforms the other three. Furthermore, we also train a model using the examples with high variability of confidence within “moderate to answer” examples, which shows further performance improvement in comparison with that trained with examples randomly sampled from “moderate to answer” examples. Future work includes applying the analysis procedure of this paper to several popular MRC benchmark datasets other than SQuAD [8] and investigating whether the similar results are obtained. We also work on deeper analysis of the characteristics of “easy to answer”, “moderate to answer”, and “hard to answer” examples to find out features that are related to the disparity of training effectiveness.

References

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

[2] K. M. Hermann, T. Kociskỳ, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. In *Proc. 28th NIPS*, pp. 1693–1701, 2015.

[3] F. Hill, A. Bordes, S. Chopra, and J. Weston. The goldilocks principle: Reading children's books with explicit memory representations. In *Proc. 4th ICLR*, pp. 1–13, 2016.

[4] D. Kaushik and Z. C. Lipton. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proc. EMNLP*, pp. 5010–5015, 2018.

[5] H. Li, T. Chen, S. Bai, T. Utsuro, and Y. Kawada. MRC examples answerable by BERT without a question are less effective in MRC model training. In *Proc. 1st AACL-IJCNLP: Student Research Workshop*, pp. 146–152, 2020.

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[7] T. Onishi, H. Wang, M. Bansal, K. Gimpel, and D. McAllester. Who did what: A large-scale person-centered cloze dataset. In *Proc. EMNLP*, pp. 2230–2235, 2016.

[8] R. Pranav, Z. Jian, L. Konstantin, and L. Percy. SQuAD: 100,000+ questions for machine comprehension of text. In *Proc. EMNLP*, pp. 2383–2392, 2016.

[9] R. Pranav, J. Robin, and L. Percy. Know what you don't know: Unanswerable questions for SQuAD. In *Proc. 56th ACL*, pp. 784–789, 2018.

[10] S. Swayamdipta, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N. A. Smith, and Y. Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proc. EMNLP*, pp. 9275–9293, 2020.

[11] J. Weston, A. Bordes, S. Chopra, A. M. Rush, B. van Merriënboer, A. Joulin, and T. Mikolov. Towards AI-complete question answering: A set of prerequisite toy tasks. In *Proc. 4th ICLR*, pp. 1–14, 2016.

[12] T. Wolf, J. Chaumond, L. Debut, V. Sanh, C. Delangue, A. Moi, P. Cistac, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers: State-of-the-art natural language processing. In *Proc. EMNLP: System Demonstrations*, pp. 38–45, 2020.