Answerable or Not: Devising a Dataset for Extending Machine Reading Comprehension

Mao Nakanishi **Tetsunori Kobayashi** **Yoshihiko Hayashi**
School of Science and Engineering, Waseda University
Waseda-machi 27, Shinjuku, Tokyo 1690042, Japan
nakanishi@pcl.cs.waseda.ac.jp koba@waseda.jp yshk.hayashi@aoni.waseda.jp

Abstract

Machine reading comprehension (MRC) has recently attracted attention in the fields of natural language processing and machine learning. One of the problematic presumptions with current MRC technologies is that each question is assumed to be answerable by looking at a given text passage. However, to realize human-like language comprehension ability, a machine should also be able to distinguish not-answerable questions (NAQs) from answerable questions. To develop this functionality, a dataset incorporating hard-to-detect NAQs is vital; however, its manual construction would be expensive. This paper proposes a dataset creation method that alters an existing MRC dataset, the Stanford Question Answering Dataset, and describes the resulting dataset. The value of this dataset is likely to increase if each NAQ in the dataset is properly classified with the difficulty of identifying it as an NAQ. This difficulty level would allow researchers to evaluate a machine's NAQ detection performance more precisely. Therefore, we propose a method for automatically assigning difficulty level labels, which basically measures the similarity between a question and the target text passage. Our NAQ detection experiments demonstrate that the resulting dataset, having difficulty level annotations, is valid and potentially useful in the development of advanced MRC models.

1 Introduction

Language understanding is one of the ultimate goals of artificial intelligence. Because the machine understanding of human language is hard to define, one often presumes that if a machine can answer a set of questions under given circumstances, then it understands language. Machine reading comprehension (MRC) developed along this direction and has recently attracted a great deal of attention in natural language processing (NLP) and machine learning.

The tasks of MRC are defined in various ways (Richardson et al., 2013; Mostafazadeh et al., 2017; Weston et al., 2015; Chen et al., 2016; Rajpurkar et al., 2016). Some tasks (Chen et al., 2016; Hill et al., 2016) provide a set of fill-in-the-blank style questions. The most prominent tasks (Rajpurkar et al., 2016), however, allow a machine to locate an answer span in given text passages. It is a highly efficient scheme, because the task can be formulated by the recognition of the start and end positions of the passage. This model facilitates the development of machine learning approaches and the automatic evaluation of performances. However, as criticized by (Jia and Liang, 2017), a machine can succeed in a task of this kind by remembering and recalling linguistic patterns that are prominent in a target dataset. This means that machine comprehension of this type would be an insufficient measure for assessing the degree of machine understanding of language.

To tackle this issue, Jia and Liang (2017) developed a dataset, in which passages were altered to include adversarial sentences that could lead to wrong answers. Contrary to their approach, which modified text passages, this paper proposes to extend MRC by incorporating questions for which relevant answers cannot be found in the given text passages. These questions are called *not-answerable questions (NAQs)*.

Proceedings of the 27th International Conference on Computational Linguistics, pages 973–983
Santa Fe, New Mexico, USA, August 20-26, 2018.

This partly meets our expectation, which can be stated as, "a good MRC system, given text passages, should properly detect and reject NAQs."

To develop an appropriate functionality, a dataset that accommodates NAQs and answerable questions is mandatory. Furthermore, each NAQ in the dataset should not be easily detected. However, the manual construction of such a dataset would be very costly. Therefore, this paper proposes an automatic dataset creation method for incorporating NAQs by making use of an existing MRC dataset. We created such a dataset by modifying a popular MRC dataset, the Stanford Question Answer Dataset (SQuAD) (Rajpurkar et al., 2016).

The value of such a dataset would increase if each NAQ in the dataset were to be graded according to the difficulty level used to determine question as not-answerable. These difficulty level annotations would allow us to choose a subset where the difficulty levels of NAQs could be adjusted as required. Therefore, we propose a method for automatically assigning difficulty level labels to NAQs. The proposed method employs a set of similarity features that are highly effective in determining the not-answerability of a question. NAQ detection experiments, employing an answerable/not-answerable binary classifier and a "null-answer detector" are conducted, showing that the resulting dataset with its respective difficulty level labels can be effectively used in the study of the genuine machine understanding of a language.

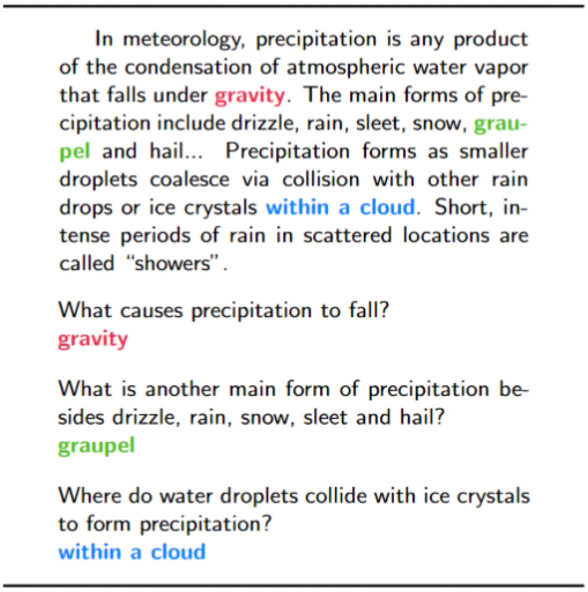
2 Related Work

Recently, several MRC datasets have been developed. The characteristics of each MRC dataset varies, depending on the purpose. For example, the Childrens Book Test dataset (Hill et al., 2016) was built from books for children, where the 21st sentence that follows the preceeding 20 sentences is employed as a "question." Additionally, the CNN/Daily dataset (Chen et al., 2016) collects news articles with bullet-pointed summaries, in which each summary is converted into a question. Among the many MRC datasets, this section introduces two datasets, WIKIQA (Yang et al., 2015) and SQuAD (Rajpurkar et al., 2016). Of these, the latter is used in this work. Additionally, we address potential problems with SQuAD.

2.1 Related MRC Datasets

WIKIQA: WIKIQA is one of the few MRC datasets that contains questions without answers, which is a main concern of this paper. WIKIQA was created to capture the characteristics of natural queries asked by people in the real world. Each question in this dataset was taken from a search engine's actual query log, and the corresponding Wikipedia summary paragraph is employed as a text passage, where each of the contained sentences is treated as an answer candidate. This means that a machine is only required to infer the positive or negative status of each sentence. This problem can be often solved by only looking at the questions; the ability of reading comprehension is not necessarily required. WIKIQA maintains 3,047 questions, of which about two thirds are NAQs, as they never have a positive sentence in the corresponding paragraphs. Moreover, 20.3% of the answers share no content words with the questions, contributing to the elevated difficulty level of the dataset. Unfortunately, this dataset is relatively small, and the amount of training data is inevitably limited.

SQuAD: SQuAD is an MRC dataset accommodating 107,785 question-and-answer (QA) pairs on 536 Wikipedia articles. A passage is associated with, at most, five QA pairs, and each question has the corresponding answer, forming a span in the associated passage. The evaluation metrics are exact match (EM) and F1. An EM score measures the percentage of predictions that match any one of the ground truth answers exactly. An F1 score is a metric that measures the average overlap between the prediction and the ground truth answer. Figure 1, adopted from (Rajpurkar et al., 2016), exemplifies three QA pairs taken from a paragraph in a Wikipedia article, whose topic is precipitation. Note that each of the three questions refers to the same passage, and the corresponding answer can be found as a span within the passage. As detailed in the next section, we create an MRC dataset with NAQs by modifying an existing dataset. If a good source dataset is properly chosen, this strategy could prevent the shortage of QAs and enjoy the advantages of the existing dataset. Among the many MRC datasets, we choose SQuAD as the source, because it is a rather large-scaled dataset of the answer-in-the-text style, which functions as a good starting point to pursue the study of genuine machine understanding of a language.



In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Figure 1: Example of QA pairs on a text passage in SQuAD.

2.2 Potential Problems with SQuAD

Jia and Liang (2017) make a point that the extent to which MRC systems truly understand language remains unclear. To reward systems that have real language understanding abilities, they proposed an adversarial evaluation method for SQuAD. Their method challenges the MRC system by altering the passages to include adversarial sentences that could lead to wrong answers. The results of their experiments proved that current MRC systems are easily affected by inserted adversarial sentences, concluding that the machines need to understand language more precisely. Their approach can be contrasted with ours, in that they modify text passages, whereas we propose to incorporate NAQs. However, both share the same objective of extending the conventional MRC framework.

Sugawara et al. (2017) evaluated a published MRC dataset for ascertaining its quality. They adopted two metrics: prerequisite skills and readability. Prerequisite skills include abilities, such as object tracking, coreference resolution, and logical reasoning. Readability was evaluated based on NLP metrics, such as average sentence length in words and adverb variation. Their evaluation analyses revealed that the prerequisite skills correlated more with dataset quality than readability. They concluded that SQuAD passages were not very readable, whereas the questions were relatively easy to answer. Although this insight does not have any direct connection with the present work, it should be noted, because it provided considerable analyses of SQuAD.

3 Dataset Creation

We propose an automatic method for creating a dataset that is useful in the development of an NAQ detection mechanism. The proposed method modifies an existing popular MRC dataset (i.e., SQuAD), which avoids the manual construction of a dataset from scratch. This section first describes the dataset creation method and then proposes a method for grading the difficulty level of an NAQ. Here the "difficulty level" signifies how difficult a question is for a machine to identify as an NAQ[1]. The description of the resulting dataset concludes this section.

[1] A set of Python scripts for creating an NAQ dataset from SQuAD is available at `https://github.com/nknsh0000/createNAQs`.

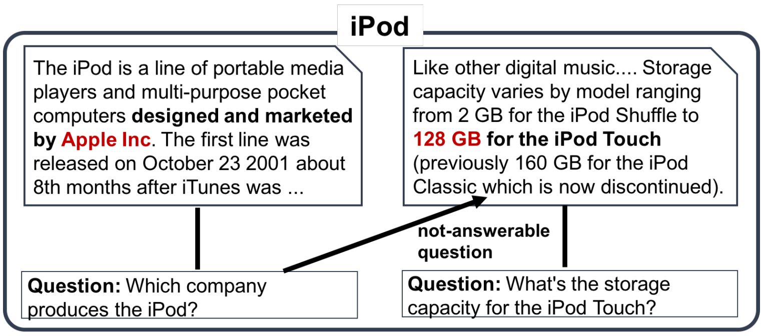

iPod
The iPod is a line of portable media players and multi-purpose pocket computers designed and marketed by Apple Inc. The first line was released on October 23 2001 about 8th months after iTunes was ...
Like other digital music.... Storage capacity varies by model ranging from 2 GB for the iPod Shuffle to 128 GB for the iPod Touch (previously 160 GB for the iPod Classic which is now discontinued).
not-answerable question
Question: Which company produces the iPod?
Question: What's the storage capacity for the iPod Touch?

Figure 2: Creation of NAQ with our proposed creation strategy.

3.1 Devising NAQs

3.1.1 The Strategy

The priority in the creation of an applicable dataset is to incorporate hard-to-detect NAQs. It seems rather easy to devise an NAQ. Randomly selected or generated questions can cater to the most fundamental requirements. However, the questions gathered by this manner could be easily detected by a machine. Thus, they are an inadequate device for assessing the degree of machine language comprehension.

The similarity between a question and the corresponding text passage plays a key role in the creation of NAQs. More specifically, an NAQ is created by simply taking a question that was originally associated with another text passage that is highly similar to the text passage of the current target. This practical method can be applied to any MRC datasets, where the answer to a question is restricted to a span in the corresponding text passage.

3.1.2 Creation of NAQs using SQuAD

Adjacent text passages in SQuAD are taken from Wikipedia articles, meaning that these passages are very likely to share a topic and thus be similar. This pattern allows us to simply select an adjacent passage as the target text passage of an NAQ. Figure 2 illustrates the procedure for NAQ creation from two pairs of a text passage and a question. The question, "which company produces the iPod?," originally associated with the left passage, is taken as an NAQ of the right passage. Notice here that the entity, "iPod," resides in the passage, potentially confusing a machine. Questions that accidentally have the corresponding answers in a shifted passage are thus removed[2].

3.2 Grading NAQ Difficulties

Although the difficulty levels of created NAQs may vary, their proper annotations can be effectively utilized. We can extract a subset of the entire dataset by choosing the difficulty levels and by enabling more detailed evaluation of a machine's NAQ detection performance. Difficulty levels should be automatically determined by referring to mechanically-extractable features, rather than relying on human assessments.

3.2.1 Feature Detection

The selection of a feature to well-estimate the difficulty level of an NAQ is not trivial. To accomplish this sub-task, we first collect potentially useful features and train an answerable/not-answerable binary classifier. We then conduct ablation tests to find the most effective feature among the candidates.

Potentially useful features can be divided into two feature groups: Individual features and Similarity features.

Individual features: To represent a passage and a question individually, an averaged word embedding vector and its TF-IDF weighted version are respectively computed for each of the passages and

[2] We removed 1,825 questions.

each of the questions. We employed pre-trained 100-dimensional GloVe (Pennington et al., 2014) word embedding vectors[3].

Similarity features: A largely irrelevant NAQ that is easy-to-detect is expected to have lower similarity to the given passage. We thus compute five types of similarity features as listed below.

- (a) max_word_sim: the maximum similarity between a word in the question $q = (\boldsymbol{w_1^q}, \ldots, \boldsymbol{w_m^q}, \ldots, \boldsymbol{w_M^q})$ and a word in the passage $p = (\boldsymbol{w_1^p}, \ldots, \boldsymbol{w_n^p}, \ldots, \boldsymbol{w_N^p})$. Thus, this similarity feature is formulated as: $max\_word\_sim = \max_{m,n} \cos(\boldsymbol{v_m^q}, \boldsymbol{v_n^p})$;
- (b) ave_word_sim: the averaged similarity between a word in the query and its most similar word in the passage: $ave\_word\_sim = \frac{1}{M} \sum_{m=1}^{M} \max_n \cos(\boldsymbol{v_m^q}, \boldsymbol{v_n^p})$;
- (c) the BLEU score (Papineni et al., 2002);
- (d) t_{cos}: the cosine similarity between each of the TF-IDF bag-of-words vectors; and
- (e) gt_{cos}: the cosine similarity between the averaged TF-IDF weighted word embedding vectors.

In the calculation of all features, we treat all out-of-vocabulary words as ”unknown”, which means that these words are simply skipped.

Binary classification: We conducted answerable/not-answerable binary classification experiments by employing the above-mentioned feature groups. In the experiments, we compared four classification algorithms (i.e., random forest (RF), linear regression (LR), support vector machine (SVM), and AdaBoost). Each were trained with 45,000 NAQs and the same number of answerable questions, randomly chosen from the created dataset. A 10-fold cross validation was applied in each run. As Table 1 clearly shows, similarity features could be utilized as a relatively good indicator.

Classifier	RF	LR	SVM	AdaBoost
Individual Features	0.597	0.563	0.565	0.554
Similarity Features	0.847	**0.852**	0.851	**0.852**

Table 1: Accuracy results of the binary classification.

Ablation tests: Based on the experimental results, we employed similarity features with the LR classifier. We then conducted ablation tests, where each similarity feature was ablated under the same experimental settings. Table 2 summarizes the results. As indicated by the greatest degradation, the feature of average similarity between words in the question and the passage, ave_word_sim, contributed the most. Given these results, we simply exploited the ave_word_sim feature as the indicator of difficulty level.

Ablated feature	max_word_sim	ave_word_sim	BLEU	t_{cos}	gt_{cos}
Accuracy	0.851	**0.812**	0.830	0.852	0.851

Table 2: Accuracy results from the ablation tests (comparing accuracy: 0.852).

3.3 Created Dataset

We finally completed our dataset by integrating the created NAQs with the original answerable questions from SQuAD. Table3 measures the number of questions in the created dataset.

Furthermore, we assigned a difficulty level label to each NAQ in the dataset by using the similarity feature, ave_word_sim, with the following criteria. Table 4 classifies the distribution of assigned difficulty levels. Note that the NAQs creation and the difficulty level assignment were individually conducted for train set and dev set.

[3] http://nlp.stanford.edu/data/glove.6B.zip

	Train		Dev	
	#passages	#questions	#passages	#questions
Answerable questions	18,896	87,599	2,067	10,570
NAQs	17,071	75,155	2,065	9,762

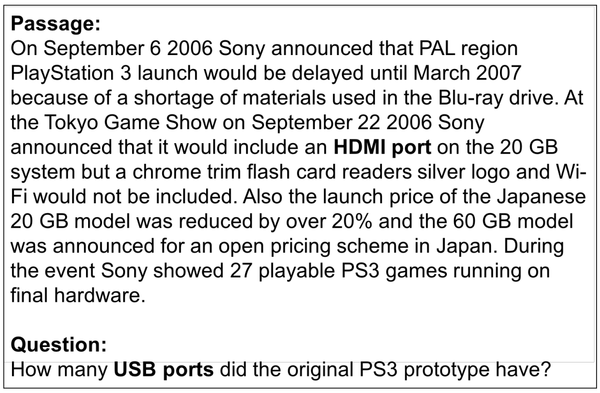
Table 3: The number of questions in the created dataset.

- LEVEL1 (easy): $0.0 \leq ave\_word\_sim < 0.5$
- LEVEL2 (moderate): $0.5 \leq ave\_word\_sim < 0.7$
- LEVEL3 (difficult): $0.7 \leq ave\_word\_sim \leq 1.0$

	Ranges of ave_word_sim	Train	Dev
ALL	$0.0 \leq ave\_word\_sim \leq 1.0$	75,155	9,762
LEVEL1	$0.0 \leq ave\_word\_sim < 0.5$	9,686	823
LEVEL2	$0.5 \leq ave\_word\_sim < 0.7$	57,730	4,636
LEVEL3	$0.7 \leq ave\_word\_sim \leq 1.0$	7,739	4,303

Table 4: Distribution of NAQ difficulty levels.

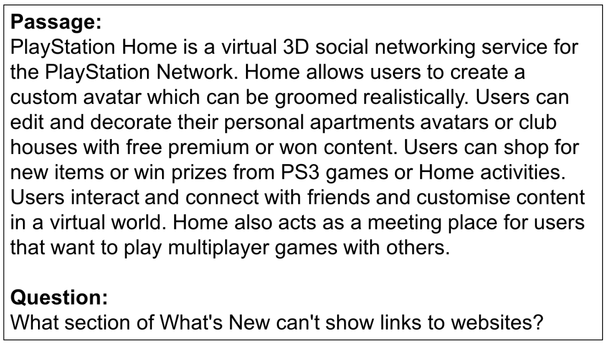
For reference, Figure 3 and Figure 4 respectively exemplify a LEVEL3 (difficult) and a LEVEL1 (easy) questions.



Passage:
On September 6 2006 Sony announced that PAL region PlayStation 3 launch would be delayed until March 2007 because of a shortage of materials used in the Blu-ray drive. At the Tokyo Game Show on September 22 2006 Sony announced that it would include an **HDMI port** on the 20 GB system but a chrome trim flash card readers silver logo and Wi-Fi would not be included. Also the launch price of the Japanese 20 GB model was reduced by over 20% and the 60 GB model was announced for an open pricing scheme in Japan. During the event Sony showed 27 playable PS3 games running on final hardware.

Question:
How many **USB ports** did the original PS3 prototype have?

Figure 3: An example of difficult NAQ.



Passage:
PlayStation Home is a virtual 3D social networking service for the PlayStation Network. Home allows users to create a custom avatar which can be groomed realistically. Users can edit and decorate their personal apartments avatars or club houses with free premium or won content. Users can shop for new items or win prizes from PS3 games or Home activities. Users interact and connect with friends and customise content in a virtual world. Home also acts as a meeting place for users that want to play multiplayer games with others.

Question:
What section of What's New can't show links to websites?

Figure 4: An example of an easy NAQ.

4 NAQ Detection Experiments and the Results

We conducted two types of NAQ detection experiments: primary experiments for comparing model architectures, and additional experiments for confirming the validity of the difficulty levels.

4.1 Comparing Model Architectures

In Section 3.2, answerable/not-answerable binary classification experiments were described. They were conducted to explore efficient features for the difficulty-level grading. However, the model architecture for NAQ detection was not necessarily limited to the described one. Rather, we adopted an existing MRC answering model for the present purpose. Namely, we exploited BiDAF (Min et al., 2016), a popular neural network model for MRC/QA in the present NAQ detection experiments.

Figure 5 overviews two BiDAF-based architectures for NAQ detection: (a) binary classification model with answer span confidences, and (b) "null-answer" question detection model.

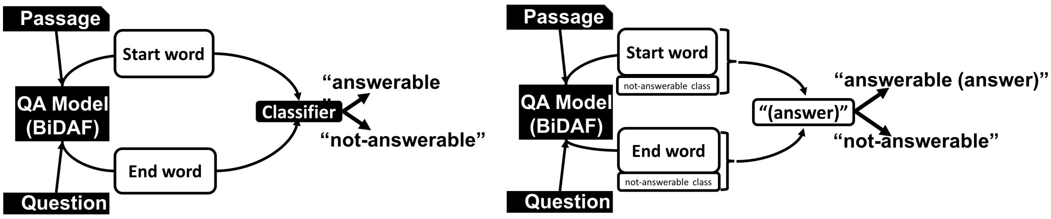


Figure 5: BiDAF-based model architectures for NAQ detection experiments: (left) binary classification model; (right) "null-answer" detection model.

(a) Binary classification model with answer span confidences: Like many existing MRC models, BiDAF estimates both start and end words of an answer candidate in the given text passage, allowing us to utilize two confidences associated with these two words as features. In this paper, we refer to the confidence of a start-word vector as s and that of an end word as e. The numbers of dimensions of s and e were aligned with the number of the words in the longest text passage. This resulted in 673 dimensions. In this experiment, 75,155 NAQs and the same number of answerable questions were randomly chosen for training, and 9,762 NAQs and the same number of answerable questions were randomly chosen for testing.

(b) "null-answer" question detection model: The BiDAF model can be leveraged another way. This model explicitly adds a "null-answer" class by properly representing an empty span associated with a null-answer. Through a preliminary experiment, we expressed it by locating both the start and end positions at the very last word of the text passage. The same set of answerable questions and NAQs was also used in this experiment.

4.1.1 The Results

Table 5 summarizes the classification accuracy with the binary classification model, where four types of classifiers (RF, LR, SVM, and AdaBoost) were particularly compared. As seen in Table 5, the highest accuracy of 0.813 was achieved with the RF classifier that employs both features, (yp_{start} and yp_{end}). These results were worse than those of the classifiers with similarity features utilized in the difficulty level grading, presented in Section 3.2.

Table 6, on the other hand, presents the classification results of accuracy with the null-answer detection model, where the accuracy was as high as 0.895, which greatly outperformed other models, including the one described in Section 3.2. Moreover, the recall rate for recovering answerable questions was as high as 0.936, whereas that for recovering NAQ was as low as 0.854, indicating that the detection of NAQs is more difficult. If we only consider the results with answerable questions, the EM score was 0.607, which was moderately worse than 0.680, achieved with the model trained without NAQs. Thus, the percentage

Classifier		RF	LR	SVM	AdaBoost
Features	yp_{start}	0.807	0.774	0.760	0.808
	yp_{end}	0.785	0.756	0.751	0.792
	yp_{start}, yp_{end}	**0.813**	0.780	0.774	0.813

Table 5: Accuracy results with the binary classification model.

	All	Answerable questions	NAQs
Accuracy of NAQ detection	**0.895**	0.936	0.854
EM scores	-	0.607	-

Table 6: Accuracy results with the null-answer detection model.

of misclassifying answerable questions as not-answerable was only 6.36% (1-0.936), whereas that of misclassifying NAQs as answerable was 14.6% (1-0.854).

With the null-answer model, it is expected that correct answers will be recovered. The error rate of answer extraction for the correctly identified answerable questions was 0.33, which is comparable with 0.32, achieved with the model trained without NAQs. This small difference supports the assumption that the model would have misclassified many answerable questions as NAQs.

4.2 Validating the Difficulty Level Grading

We can extract a subset of the entire dataset by choosing the difficulty levels, which enables more detailed evaluation of the machine's NAQ detection performance. If the detection accuracies degrade with the increase of difficulty levels, then the assigned difficulty levels are valid. Moreover, if the NAQ detection accuracies are low enough, the created dataset is useful for the dev of the functionality to distinguish NAQs from answerable questions.

For each class of the difficulty level, we randomly extracted the same number of answerable questions from the SQuAD dataset as the NAQs from our dataset, and conducted the binary classification for the dev set. Again, we compared the binary classification model (with the RF classifier) with the null-answer detection model.

4.2.1 The Results

The results in Table 7 show that the accuracies degrade with the increase of the difficulty level, indicating that the difficulty grading is valid. Additionally, we confirmed that the null-answer model was superior to the present binary classification model. Moreover, the results for the most difficult level NAQs gave significantly lower accuracy figures, signaling that the models for detecting NAQs still have room for improvement. Nevertheless, we can construct a sub-dataset with designated difficulty levels from the entire dataset, which could be effectively exploited in the development of MRC models that deal with NAQs.

4.3 Error Analysis

Errors are divided into two cases, where the machine: (a) judged an NAQ as answerable (false answerable), or (b) judged an answerable question as an NAQ (false not-answerable).

Model	Binary classification	Null-answer detection
LEVEL1	0.860	0.953
LEVEL2	0.855	0.948
LEVEL3	0.748	0.734

Table 7: Accuracy results of NAQ detection experiments for each difficulty level.

4.3.1 False answerable

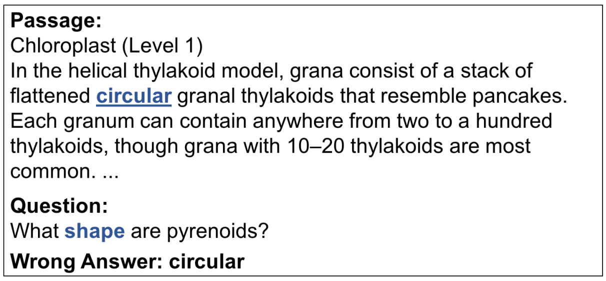
Figures 6, 7 and 8 respectively display a false answerable error case, where the machine wrongly predicted a question as answerable in each difficulty level.

In the Level-1 question exemplified in Fig. 6, the machine wrongly answered the shape of thylakoid rather than that of pyrenoids. This error could be attributed to the existence of the keyword ”shape” appeared in the question. The passage actually had nothing to do with pyrenoids. In the example shown in Fig. 7, the passage contains the sentence, ”In 1893, Tesla returned to his birthtown, Smijian.” The machine's prediction would have been strongly affected by the existence of not only ”Tesla” but ”go”: ”go” in the question and ”return” in the passage exhibit a level of semantic similarity. In this example, the machine should have recognize that ”Smiljan” and ”Karlovac” are totally different placements, whose similarity would not be taken into account. Finally, Fig. 8 displays a Level-3 question. Similar to the example given in Fig. 7, the machine was again not able to recognize the difference in nouns; in this case the focused nouns are compounds, each of them designates a specific event.

These examples suggest that the machine estimated the NAQs as answerable questions especially when some words overlapped between passage and question. In these cases, the machine tends to simply predict a span whose estimated semantic type is matched with the type of the interrogative.

4.3.2 False not-answerable

In false not-answerable cases, we observed differences in the mean length of questions and that of passage as counted in Table 8. This table suggests that the machine tended to recognize answerable questions as NAQs when the passage length is short or when the ground truth answer length is long.

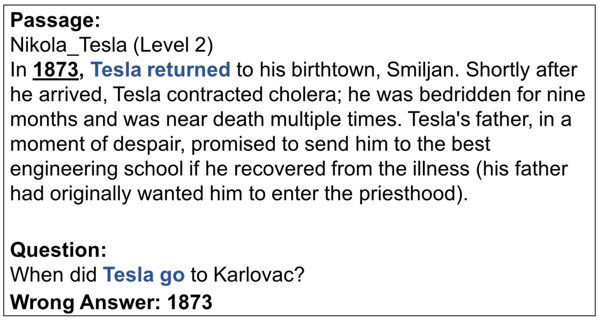


Passage:
Chloroplast (Level 1)
In the helical thylakoid model, grana consist of a stack of flattened **circular** granal thylakoids that resemble pancakes. Each granum can contain anywhere from two to a hundred thylakoids, though grana with 10–20 thylakoids are most common. ...

Question:
What **shape** are pyrenoids?

Wrong Answer: circular

Figure 6: A Level-1 NAQ that the machine misjudged as an answerable question.



Passage:
Nikola_Tesla (Level 2)
In **1873, Tesla returned** to his birthtown, Smiljan. Shortly after he arrived, Tesla contracted cholera; he was bedridden for nine months and was near death multiple times. Tesla's father, in a moment of despair, promised to send him to the best engineering school if he recovered from the illness (his father had originally wanted him to enter the priesthood).

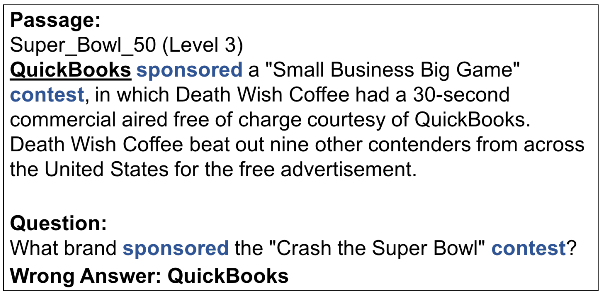
Question:
When did **Tesla go** to Karlovac?

Wrong Answer: 1873

Figure 7: A Level-2 NAQ that the machine misjudged as an answerable question.

5 Concluding Remarks

Machine understanding of language is difficult to define and accomplish. However, we must approach this issue by developing a computational mechanism, along with the relevant resources. As a step along this direction, in this paper, an automatic dataset creation method with which NAQs were incorporated was proposed. We created a dataset by altering an existing QA dataset, SQuAD.



Passage:
Super_Bowl_50 (Level 3)
QuickBooks **sponsored** a "Small Business Big Game" **contest**, in which Death Wish Coffee had a 30-second commercial aired free of charge courtesy of QuickBooks. Death Wish Coffee beat out nine other contenders from across the United States for the free advertisement.

Question:
What brand **sponsored** the "Crash the Super Bowl" **contest**?
Wrong Answer: QuickBooks

Figure 8: A Level-3 NAQ that the machine mistook as an answerable question.

	Answerable questions	
	Correctly predicted as answerable	Incorrectly predicted as NAQ
Passage	777.39	722.01
Answer	2.82	3.25
Answer/Passage Ratio [%]	0.42	0.53

Table 8: The lengths of passages and answers contained answerable questions.

One of the key requirements for such a dataset would be the acquisition of NAQs that are difficult to identify as not-answerable. Given this sub-goal, a method for assigning the difficulty levels to NAQs was also described. The results of NAQ detection experiments demonstrated that the assigned difficulty levels were valid, because the detection accuracies were degraded with the increase in difficulty level. This means that we acquired a way to examine and evaluate NAQ detection methods and models more precisely. Moreover, the lower accuracy results confirmed that there is still scope for improvement in the NAQ detection models. Therefore, the created dataset could be effectively utilized in developing the NAQ detection functionality.

For future work, we should develop a more difficult MRC dataset by further investigating into the notion of the difficulty in NAQ detection, which would inherently require us to incorporate various aspects of human language understanding. Along with this direction, we will devise a set of questions, for which confusing items reside in the target text passages.

References

Chen, D., Bolton, J., and Manning, C. D. (2016). A thorough examination of the CNN/Daily Mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2367, Berlin, Germany.

Hill, F., Bordes, A., Chopra, S., and Weston, J. (2016). The goldilocks principle: Reading children's books with explicit memory representations. In *Proceedings of the 4th International Conference on Learning Representations*, San Juan, Puerto Rico.

Jia, R. and Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark.

Min, S. J., Aniruddha, K., Ali, F., and Hannaneh, H. (2016). Bidirectional attention flow for machine comprehension. *CoRR*, abs/1611.01603.

Mostafazadeh, N., Roth, M., Louis, A., Chambers, N., and Allen, J. (2017). Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51, Valencia, Spain. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Stroudsburg, PA, USA.

Pennington, J., Socher, R., and Manning, D. C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, USA.

Richardson, M., Burges, J. C. C., and Renshaw, E. (2013). Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Emprical Methods in Natural Language Processing*, pages 193–203, USA.

Sugawara, S., Kido, Y., Yokono, H., and Aizawa, A. (2017). Evaluation metrics for machine reading comprehension: Prerequisite skills and readability. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 806–817, Vancouver, Canada.

Weston, J., Bordes, A., Chopra, S., and Mikolov, T. (2015). Towards ai-complete question answering: A set of prerequisite toy tasks. *CoRR*, abs/1502.05698.

Yang, Y., Yih, S. W., and Meek, C. (2015). 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.