NQuAD: 70,000+ Questions for Machine Comprehension of the Numerals in Text

Chung-Chi Chen
Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
cjchen@nlg.csie.ntu.edu.tw

Hen-Hsen Huang
Institute of Information Science, Academia Sinica, Taiwan
MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
hhhuang@iis.sinica.edu.tw

Hsin-Hsi Chen
Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
hhchen@ntu.edu.tw

ABSTRACT

Numeral information plays an important role in narratives of several domains such as medicine, engineering, and finance. Previous works focus on the foundation exploration toward numeracy and show that fine-grained numeracy is a challenging task. In machine reading comprehension, our statistics show that only a few numeral-related questions appear in previous datasets. It indicates that few benchmark datasets are designed for numeracy learning. In this paper, we present a Numeral-related Question Answering Dataset, NQuAD, for fine-grained numeracy, and propose several baselines for future works. We compare NQuAD with three machine reading comprehension datasets and show that NQuAD is more challenging than the numeral-related questions in other datasets. NQuAD is published under the CC BY-NC-SA 4.0 license for academic purposes.

CCS CONCEPTS

• **Computing methodologies** → **Language resources**.

KEYWORDS

Numeracy, machine reading comprehension, cloze test

ACM Reference Format:
Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. NQuAD: 70,000+ Questions for Machine Comprehension of the Numerals in Text. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3459637.3482155

1 INTRODUCTION

Recently, numerals in the table [14] and the content [16] of a document have attracted more researchers' attention. Machine numeral understanding is one of the emerging research topics and is still in the early-stage. Naik et al. [18] and Wallace et al. [27] probe the numeracy of word embeddings. Spithourakis and Riedel [26] evaluate the numeracy of the language models. The experimental results of the previous works indicate that neural network models tend to get confused by closer numerals [5, 18]. A dataset designed specially for the fine-grained numeracy analysis is mandatory. In this paper, we create a **N**umeral-related **Qu**estion **A**nswering **D**ataset, named NQuAD, by selecting the fine-grained numeral options from news articles and asking the machine to predict the correct the option.

News Article:
Major banks take the lead in self-discipline. The five major banks' newly-imposed mortgage interest rates climbed to **1.986%** in May. ... Also approaching **2%** integer alert ... Up to **2.5%** ... Also increased by **0.04** percentage points from the previous month ... Prevent the housing market bubble from fully starting.

Question Stem: Driven by self-discipline, the five major banks' new mortgage interest rates are approaching nearly _____%.

Answer Options: (A) 0.04 (B) 1.986 (C) 2 (D) 2.5

Answer: (C)

Figure 1: An example question in NQuAD.

The questions satisfy one of the following conditions are considered as a numeral-related question: (1) there exists at least one numeral in the answer snippets for the machine reading comprehension (MRC) questions; and (2) all answer options contain at least one numeral for the multiple-choice questions. Figure 1 shows an example in our dataset, including a news article, a question stem and four answer options. Based on the above definition, Table 1 shows the statistics of three Chinese MRC datasets (CMRC-2017 [9], DRCD [24], and CMRC-2018 [8]) and four English multiple-choice MRC datasets (MCTest [22], RACE [15], MCScript [19], and ARC [7]). We find that only a few numeral-related instances are collected in these datasets. Such a finding supports that the proposed dataset is distinctive.

Additionally, all of the previous works examine the numeracy on English dataset [5, 18, 26, 27]. This is the first work that presents the results of numeracy in Chinese data. We compare NQuAD with the previous MRC datasets in terms of difficulty and propose a

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3482155

Table 1: Comparison with the other MRC datasets.

Chinese	NQuAD	CMRC-2017	DRCD	CMRC-2018
# Questions	71,998	5,000	33,953	14,363
# Num-related	71,998	0	6,575	3,506
% of Num-related	100.00%	0.00%	19.37%	24.41%
English	MCTest	RACE	MCScript	ARC
# Questions	2,640	97,687	13,939	7,787
# Num-related	46	3,343	249	233
% of Num-related	1.74%	3.42%	1.79%	2.99%

Numeracy-**E**nhanced **Mo**del (NEMo) for the proposed task. Our experimental results also indicate that adding a numeracy encoder into models can significantly improve the performance.

2 RELATED WORK

Recently, testing numeracy attracts many researchers' attentions. Spithourakis and Riedel [26] use the root mean squared error to evaluate the prediction of the language models. They show that the performance of the best model in the clinical dataset is 989.84. Chen et al. [5] propose a dataset, named Numeracy-600K. In Numeracy-600K, the task is to predict the magnitude in the blank of a market comment and an online article title. They show that machine can achieve micro-averaged F1-score of 80% in this task, and indicate that the models perform worse when the difference between the exaggerated numeral and the true numeral is small. All of the previous works find that the machine performs well for the numerals having significant differences, but performs worse for the close numerals. That is what inspires us to present a dataset for test fine-grained numeracy. In this paper, we construct NQuAD by selecting the top 4 closest candidates from the related articles as the options for the question stem. Total 87.10% of the questions in the proposed dataset tally with the condition that the average difference between the options and the answer is less than 10. That shows our dataset is more suitable for the fine-grained numeracy test.

Wallace et al. [27] indicate that the character-level recurrent neural network performs well in learning numeracy. Chen et al. [1] show that adding the magnitude embedding to represent the position of the digit in the numeral can provide a significant improvement for the numeral-related task. Inspired by these works, we represent a numeral with both character and magnitude embedding in the proposed model.

3 DATASET

3.1 Background of Task Setting

When mentioning numeracy or numeral-related questions, mathematical reasoning [23] or mathematical problem solving [13] comes into most people's minds. Some explorations have been done in current works [11, 20]. In the SQuAD 2.0 [21] and DROP [11], 13.34% and 68.83% of questions are numeral-related, respectively. However, unlike these works focusing on learning to answer a question in the human reading comprehension test, we pay attention to a basic but important issue: selecting a proper numeral based on the given text. This issue is important because, based on our observation, about 93% of the numerals in news headlines just copy, paraphrase, or round certain numeral in the article. That means when we attempt to generate a headline containing numeral information [6], selecting the correct numeral is an important step. Thus, in this paper, we present a pilot dataset and explore it to address the proposed issue.

Table 2: Statistics of the collected news.

	Headline		Content	
	#	%	#	%
Has Num.	45,075	59.74	75,296	99.80
No Num.	30,373	40.26	152	0.20

Table 3: The degree of difficulty. Note that lower means more difficult. The bold results are the best performance.

Algorithm	NQuAD	CMRC-2018	DRCD
MLIPNS	19.65%	33.41%	**74.22**%
Cosine similarity	40.26%	56.32%	30.41%
Overleap coefficient	45.09%	**59.17**%	33.02%
Monge-Elkan	**48.28**%	58.84%	35.46%

3.2 Dataset Construction

We collect news articles from the data vendor, MoneyDJ[1], and get the news articles within the period from June 22, 2013 to June 20, 2018. A total of 75,448 Chinese news articles are collected. Table 2 shows the statistics of the collected news. In this collection, 59.74% of news headlines contain at least one numeral, and 99.80% of news contents contain numerals.

Because the news articles are written by professional journalists, the qualities of both headlines and articles used in the proposed dataset are satisfiable. The contents are also trustable. Thus, we use the numerals in headlines as the ground truths of the proposed dataset. The procedure to construct NQuAD is specified as follows:

(1) We filter out those news articles that match at least one of following two conditions: (1) no numeral in the headline, and (2) less than 4 numerals in the content.
(2) We use a numeral in the headline as the target numeral, mask the target numeral from the headline, and regard the masked headline as the stem of a question. Note that we only mask one target numeral for each question. For a headline containing k numerals, we will form k questions, each of them corresponding to each individual numeral.
(3) From the news content, we select 4 distinct numerals whose values are closest to the value of the target numeral as the plausible options of the question.

A total of 43,787 news articles are selected, and 46.97% of the headlines contain more than one numeral. The average number of numerals in the headline and in the content are 1.65 and 29.48, respectively. Each numeral in each headline is used to form a question, thus we finally obtain 71,998 questions. We separate 80% of the instances as the training set and the rest of the instances form a test set.

[1] https://www.moneydj.com/

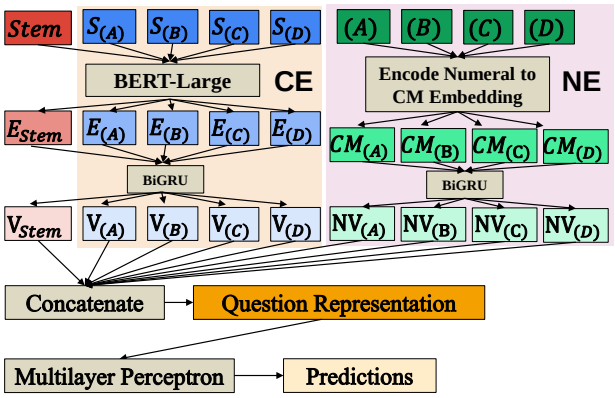

Figure 2: Architecture of the proposed model, *NEMo*.

3.3 The Degree of Difficulty

Because of different task settings between our dataset and the other Chinese datasets, we cannot employ the supervised models to all datasets directly. To compare the difficulty degree, we follow Spithourakis and Riedel [26] to use similarity as the criterion. That is, the more questions can be answered by selecting the most similar sentence in the answer options, the easier the dataset is. Here, we calculate the similarity of the question stem and the sentence containing the numeral in the answer option. Unlike previous works that reply on only one metric, we explore several similarity algorithms[2]. Table 3 shows the accuracy of the experimental results of NQuAD and other datasets. About 59.17% and 74.22% of the numeral-related questions can be answered just by literal similarity in CMRC-2018 and DRCD, respectively. In our dataset, only 48.28% of the questions can be answered by selecting the most similar option. These results support that the numeral-related questions in the proposed dataset are more difficult than those in other publically available datasets.

4 EXPERIMENTS

4.1 Methods

Figure 1 provides an example for the input of models. Figure 2 shows the architecture of the proposed **N**umeracy-**E**nhanced **Mo**del (NEMo). This model is separated into two parts, including (1) the context encoder (CE) for question stem and the sentences containing the numeral in the answer options ($S_{(A)}$, $S_{(B)}$, $S_{(C)}$, and $S_{(D)}$), and (2) the numeral encoder (NE) for the answer options ((A), (B), (C), (D)). For example, the $S_{(A)}$, $S_{(B)}$, $S_{(C)}$, and $S_{(D)}$ in Figure 1 are shown as follows.

- $S_{(A)}$: Also increased by **0.04** percentage points from the previous month
- $S_{(B)}$: The five major banks' newly-imposed mortgage interest rates climbed to **1.986%** in May.
- $S_{(C)}$: Also approaching **2%** integer alert
- $S_{(D)}$: Up to **2.5%**

[2]We adopt textdistance toolkit to calculate the similarity. Please refer to the page of the toolkit for the details of the algorithms. https://pypi.org/project/textdistance/

Table 4: Experimental results. * denotes results that are significantly different from the second-best model (BERT-BiGRU) under McNemar's test with $p < 0.05$.

Model	Accuracy
BERT Embedding Similarity	57.30%
Vanilla BERT	66.41%
BERT-BiGRU	67.15%
BERT-CNN	63.92%
NEMo	**69.95%***

In order to encode the question stem and the sentences containing the numeral in the answer options ($S_{(A)}$, $S_{(B)}$, $S_{(C)}$, and $S_{(D)}$), we adopt BERT-Large [10] as the text encoder. After encoding by BERT, we can get the tokens embeddings ($E_{(Stem)}$, $E_{(A)}$, $E_{(B)}$, $E_{(C)}$, and $E_{(D)}$). These embeddings are further put into the BiGRU model, and we can get the vectors ($V_{(Stem)}$, $V_{(A)}$, $V_{(B)}$, $V_{(C)}$, and $V_{(D)}$) for the question stem.

In the numeral encoder, we encode the numeral to the character and magnitude (CM) embeddings. That is, we use the character-level representation for the numerals and add the magnitude information to increase the numeracy of the models. A digit (0 to 9 and the standard decimal separator) in the numeral is represented by an 11-dimension one-hot vector. Because the max-length numeral in the training set is 10, the magnitude embedding is represented by a 10-dimension one-hot vector. We concatenate the digit representation with the magnitude embedding and use left-padding to fix the dimension of all numerals. Thus, a numeral is represented as a 10×11 tensor. We further put the CM embeddings to BiGRU model, and get the vectors ($NV_{(A)}$, $NV_{(B)}$, $NV_{(C)}$, and $NV_{(D)}$) for the numerals in answer options.

After encoding all information, we concatenate all vectors as the question representation and deliver it to the multilayer perceptron for predicting the answer. We compare the performance of the proposed model with those of the following methods.

- **BERT Embedding Similarity**: We calculate the cosine similarity of the sum of tokens embeddings ($E_{(Stem)}$) and that of $E_{(A)}$, $E_{(B)}$, $E_{(C)}$, and $E_{(D)}$, and select the most similar one as the answer.
- **Vanilla BERT** [10]: This model encodes the question stem and the sentences containing the numeral in the answer options by BERT-Large and make a prediction with the multilayer perceptron.
- **BERT-BiGRU**: In this model, we remove the numeral encoder, i.e., NE, from the proposed model to show the usefulness of the NE.
- **BERT-CNN**: We change BiGRU in the BERT-BiGRU model to CNN to show the difference between different neural network models.

4.2 Results

Table 4 shows the experimental results. NEMo is significantly better than the baseline models. The results of BERT-BiGRU and the NEMo indicate that adding the proposed numeral encoder to the model is helpful for improving the performance on the numeral-related

Table 5: Ablation analysis.

Model	Accuracy
NEMo	**69.95**%
-Numeral Encoder	67.15%
-Context Encoder	55.56%

questions. The results of BERT-BiGRU and BERT-CNN show that using recurrent neural networks is better than using convolutional neural networks in the proposed task. By comparing the results in Table 3 and the performance of BERT Embedding Similarity, we find that using the embedding of the pre-trained language model to calculate the similarity between the question stem and the sentences containing the numerals in the option performs better than using the unsupervised similarity algorithms. Besides, using BERT-BiGRU can get better performance than using the Vanilla BERT model in the proposed task.

5 DISCUSSIONS

5.1 Comparison with Numeracy-600K

We compare the proposed dataset with the recent numeracy dataset, Numeracy-600K [5]. In Numeracy-600K, they aim to predict the magnitude in the blank of a market comment and an online article title. This is a coarse-grained setting because the answer candidates for all questions are the same. They show that models can achieve an 80% of micro-averaged F1-score.

In contrast, the answer candidates of instances in the proposed NQuAD depend on news articles, which can be considered a test for fine-grained numeracy. In order to provide a baseline under the vanilla task setting as previous work, i.e., only input the question stem and the candidate numerals to models, we perform ablation analysis of the proposed model. Table 5 shows the experimental results after removing the numeral encoder and removing the context encoder in NEMo. Without the information of those sentences containing the numeral in the answer options ($S_{(A)}$, $S_{(B)}$, $S_{(C)}$, and $S_{(D)}$), the performance will drop significantly. That indicates the importance of the context toward numeracy testing, and also shows that the proposed task setting and dataset are more rational for testing the numeracy of models than those of the previous work.

Additionally, several previous works [18, 26] show the difficulty of discriminating the fine-grained numeral information. In the proposed dataset, the difference between the options and the answer is quite close. As shown in Table 6, 87.10% of the differences between the options and the answer are less than 10. That provides the reason why the proposed dataset is more difficult than previous datasets.

5.2 Possible Extensions

In this section, we point out some issues of the uses of the proposed dataset and discuss some possible extensions in the future work. Different from most previous works using question answering style settings, the question type of NQuAD is cloze-style. In other words, most pre-trained language models focus on predicting masked words instead of predicting masked numerals. One of the possible solutions for this issue is that we can adopt question generation methods [12, 25] to reform the questions in the proposed dataset. For example, the question stem in Figure 1 can be transferred to "how much the new mortgage rate is?" With the auto-generated questions, many question answering methods can be explored in the proposed dataset. The question reformulation will provide a new direction to answer whether the question type influences models' performances.

Table 6: Average difference (Avg Diff) in the NQuAD. Cum. denotes cumulated.

Avg Diff (d)	# Questions	%	Cum. %
$0 \leq d < 1$	25,870	35.93	35.93
$1 \leq d < 2$	19,546	27.15	63.08
$2 \leq d < 4$	10,639	14.78	77.86
$4 \leq d < 6$	3,670	5.10	82.95
$6 \leq d < 8$	1,860	2.58	85.54
$8 \leq d < 10$	1,462	2.03	87.10
$d \geq 10$	9,287	12.90	100.00

On the other hand, in this work, we focus on the sentences that contain the answer options, i.e., $S_{(A)}$, $S_{(B)}$, $S_{(C)}$, and $S_{(D)}$, and only encode these sentences in models. When analyzing errors, we find that other sentences sometimes provide extra information for answering the question. With the full news article in the proposed dataset, future work can probe more complex task settings. For example, only the news articles and question stems are given in the proposed dataset without showing the options.

Finally, the proposed model encodes the numeral independently instead of co-training numeracy during the pretraining process. Based on the experiences of the recent studies [6, 10, 17], pretraining with related tasks or datasets can improve the ability to capture context information. Future works could explore different pretraining schemes in the proposed dataset. For instance, predicting the value [26] or the magnitude [5] could be a good direction for exploring numeral-aware pretraining settings. Additionally, understanding the category of numerals [4] is also shown to be useful in numeral-related tasks [2, 3].

6 CONCLUSION

In this work, we present a new numeral-related machine reading comprehension dataset, NQuAD.[3] We provide a pilot exploration in the proposed dataset, and propose a tailor-made model, NEMo, which enhances the numeracy of the model, for solving the numeral-related questions. We show that the numeral-related questions in NQuAD are more difficult than those in other datasets. The experimental results also show the proposed model performs better than other baseline models. In the future, we plan to explore more complex numeral-related tasks such as cross-document numeral alignment and domain-specific numeral comprehension.

ACKNOWLEDGMENTS

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST 109-2218-E-009-014, MOST 110-2634-F-002-028, and MOST 110-2221-E-002 -128 -MY3.

[3] http://nquad.nlpfin.com

REFERENCES

[1] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2019. Numeral Attachment with Auxiliary Tasks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1161–1164.

[2] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. NumClaim: Investor's Fine-grained Claim Detection. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1973–1976.

[3] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. Distilling Numeral Information for Volatility Forecasting. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*.

[4] Chung-Chi Chen, Hen-Hsen Huang, Yow-Ting Shiue, and Hsin-Hsi Chen. 2018. Numeral understanding in financial tweets for fine-grained crowd-based forecasting. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 136–143.

[5] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2019. Numeracy-600K: Learning Numeracy for Detecting Exaggerated Information in Market Comments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

[6] Jui Chu, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. Learning to generate correct numeric values in news headlines. In *Companion Proceedings of the Web Conference 2020*. 17–18.

[7] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457* (2018).

[8] Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A Span-Extraction Dataset for Chinese Machine Reading Comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 5886–5891.

[9] Yiming Cui, Ting Liu, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2018. Dataset for the First Evaluation on Chinese Machine Reading Comprehension. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. 4171–4186.

[11] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 2368–2378. https://doi.org/10.18653/v1/N19-1246

[12] Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question Generation for Question Answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 866–874. https://doi.org/10.18653/v1/D17-1090

[13] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874* (2021).

[14] Yusra Ibrahim, Mirek Riedewald, Gerhard Weikum, and Demetrios Zeinalipour-Yazti. 2019. Bridging Quantities in Tables and Text. In *2019 IEEE 35th International Conference on Data Engineering*. 1010–1021.

[15] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

[16] Matthew Lamm, Arun Chaganty, Christopher D. Manning, Dan Jurafsky, and Percy Liang. 2018. Textual Analogy Parsing: What's Shared and What's Compared among Analogous Facts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 82–92.

[17] Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, Christian Bessiere (Ed.). International Joint Conferences on Artificial Intelligence Organization, 4513–4519. Special Track on AI in FinTech.

[18] Aakanksha Naik, Abhilasha Ravichander, Carolyn Rose, and Eduard Hovy. 2019. Exploring Numeracy in Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

[19] Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal. 2018. MCScript: A Novel Dataset for Assessing Machine Comprehension Using Script Knowledge. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

[20] Shuai Peng, Ke Yuan, Liangcai Gao, and Zhi Tang. 2021. MathBERT: A Pre-Trained Model for Mathematical Formula Understanding. *arXiv preprint arXiv:2105.00377* (2021).

[21] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Melbourne, Australia, 784–789. https://doi.org/10.18653/v1/P18-2124

[22] Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 193–203.

[23] David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. 2018. Analysing Mathematical Reasoning Abilities of Neural Models. In *International Conference on Learning Representations*.

[24] Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2018. Drcd: a chinese machine reading comprehension dataset. *arXiv preprint arXiv:1806.00920* (2018).

[25] Linfeng Song, Zhiguo Wang, Wael Hamza, Yue Zhang, and Daniel Gildea. 2018. Leveraging Context Information for Natural Question Generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 569–574. https://doi.org/10.18653/v1/N18-2090

[26] Georgios Spithourakis and Sebastian Riedel. 2018. Numeracy for Language Models: Evaluating and Improving their Ability to Predict Numbers. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 2104–2115.

[27] Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP Models Know Numbers? Probing Numeracy in Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 5310–5318.