A Span-Extraction Dataset for Chinese Machine Reading Comprehension

Yiming Cui[†‡], **Ting Liu**[†], **Wanxiang Che**[†],
Li Xiao[‡], **Zhipeng Chen**[‡], **Wentao Ma**[‡§], **Shijin Wang**[‡§], **Guoping Hu**[‡]
[†]Research Center for Social Computing and Information Retrieval (SCIR),
Harbin Institute of Technology, Harbin, China
[‡]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, China
[§]iFLYTEK AI Research (Hebei), Langfang, China
[†]{ymcui,tliu,car}@ir.hit.edu.cn
[‡§]{ymcui,lixiao3,zpchen,wtma,sjwang3,gphu}@iflytek.com

Abstract

Machine Reading Comprehension (MRC) has become enormously popular recently and has attracted a lot of attention. However, the existing reading comprehension datasets are mostly in English. In this paper, we introduce a Span-Extraction dataset for Chinese machine reading comprehension to add language diversities in this area. The dataset is composed by near 20,000 real questions annotated on Wikipedia paragraphs by human experts. We also annotated a challenge set which contains the questions that need comprehensive understanding and multi-sentence inference throughout the context. We present several baseline systems as well as anonymous submissions for demonstrating the difficulties in this dataset. With the release of the dataset, we hosted the Second Evaluation Workshop on Chinese Machine Reading Comprehension (CMRC 2018). We hope the release of the dataset could further accelerate the Chinese machine reading comprehension research.[1]

1 Introduction

To read and comprehend natural languages is the key to achieve advanced artificial intelligence. Machine Reading Comprehension (MRC) aims to comprehend the context of given articles and answer the questions based on them. Various types of machine reading comprehension datasets have been proposed, such as cloze-style reading comprehension (Hermann et al., 2015; Hill et al., 2015; Cui et al., 2016), span-extraction reading comprehension (Rajpurkar et al., 2016; Trischler et al., 2016), open-domain reading comprehension (Nguyen et al., 2016; He et al., 2017), reading comprehension with multiple-choice (Richardson et al., 2013; Lai et al., 2017), etc. Along with the development of the reading comprehension dataset, various neural network approaches have been proposed and made a big advancement in this area (Kadlec et al., 2016; Cui et al., 2017; Dhingra et al., 2017; Wang and Jiang, 2016; Xiong et al., 2016; Liu et al., 2017; Wang et al., 2017; Hu et al., 2018; Wang et al., 2018; Yu et al., 2018).

We also have seen various efforts on the construction of Chinese machine reading comprehension datasets. In cloze-style reading comprehension, Cui et al. (2016) proposed a Chinese cloze-style reading comprehension dataset: People's Daily & Children's Fairy Tale. To add difficulties to the dataset, along with the automatically generated evaluation sets (development and test), they also release a human-annotated evaluation set. Later, Cui et al. (2018) propose another dataset, which is gathered from children's reading material. To add more diversity and for further investigation on transfer learning, they also provide another evaluation dataset, which is also annotated by human experts, but the query is more natural than the cloze type. The dataset was used in the first evaluation workshop on Chinese machine reading comprehension (CMRC 2017). In open-domain reading comprehension, He et al. (2017) propose a large-scale open-domain Chinese machine reading comprehension dataset (DuReader), which contains 200k queries annotated from the user query logs on the search engine. Shao et al. (2018) proposed a reading comprehension dataset in Traditional Chinese.

Though we have seen that the current machine learning approaches have surpassed the human performance on the SQuAD dataset (Rajpurkar et al., 2016), we wonder if these state-of-the-art models could also give a similar performance on the dataset of different languages. To further accelerate the development of the machine reading comprehension research, we propose a span-

[1]Resources are available: https://github.com/ymcui/cmrc2018.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 5883–5889,
Hong Kong, China, November 3–7, 2019.

[Passage] 《黄色脸孔》是柯南·道尔所著的福尔摩斯探案的56个短篇故事之一，收录于《福尔摩斯回忆录》。孟罗先生素来与妻子恩爱，但自从最近邻居新入伙后，孟罗太太则变得很奇怪，曾经凌晨时份外出，又藉丈夫不在家时偷偷走到邻居家中。于是孟罗先生向福尔摩斯求助，福尔摩斯听毕孟罗先生的故事后，认为孟罗太太被来自美国的前夫勒索，所以不敢向孟罗先生说出真相，所以吩咐孟罗先生，如果太太再次走到邻居家时，即时联络他，他会第一时间赶到。孟罗太太又走到邻居家，福尔摩斯陪同孟罗先生冲入，却发现**邻居家中的人是孟罗太太与前夫生的女儿，因为孟罗太太的前夫是黑人，她怕孟罗先生嫌弃混血儿，所以不敢说出真相。**	**[Passage]** "The Adventure of the Yellow Face", one of the 56 short Sherlock Holmes stories written by Sir Arthur Conan Doyle, is the third tale from The Memoirs of Sherlock Holmes. Mr. Munro has always been loved by his wife, but since the new neighbors recently joined, Mrs. Munro has become very strange. She used to go out in the early hours of the morning and secretly went to her neighbors when her husband was not at home. ... Mrs. Munro went to the neighbor's house again, and Holmes accompanied Mr. Munro to rush in, only to find that the neighbor's family was the daughter of Mrs. Munro and her ex-husband, **because Mrs. Munro's ex-husband was black, and she was afraid of Mr. Munro hate the mixed-race, so she did not dare to tell the truth.**
[Question] 孟罗太太为什么在邻居新入伙后变得很奇怪? **[Answer 1]** 邻居家中的人是孟罗太太与前夫生的女儿，因为孟罗太太的前夫是黑人，她怕孟罗先生嫌弃混血儿 **[Answer 2]** 邻居家中的人是孟罗太太与前夫生的女儿，因为孟罗太太的前夫是黑人，她怕孟罗先生嫌弃混血儿，所以不敢说出真相。 **[Answer 3]** 邻居家中的人是孟罗太太与前夫生的女儿，因为孟罗太太的前夫是黑人，她怕孟罗先生嫌弃混血儿，所以不敢说出真相。	**[Question]** Why Mrs. Munro became strange after the new neighbors moved in? **[Answer 1]** because Mrs. Munro's ex-husband was black, and she was afraid of Mr. Munro hate the mixed-race **[Answer 2]** because Mrs. Munro's ex-husband was black, and she was afraid of Mr. Munro hate the mixed-race **[Answer 3]** because Mrs. Munro's ex-husband was black, and she was afraid of Mr. Munro hate the mixed-race, so she did not dare to tell the truth.

Figure 1: An example of the proposed CMRC 2018 dataset (challenge set). English translation is also given for comparison.

extraction dataset for Chinese machine reading comprehension. Figure 1 shows an example of the proposed dataset. The main contributions of our work can be concluded as follows.

- We propose a Chinese span-extraction reading comprehension dataset which contains near 20,000 human-annotated questions, to add linguistic diversity in reading comprehension field.
- To thoroughly test the ability of the MRC systems, besides the development and test set, we also make a challenge set which contains carefully annotated questions that require various clues in the passage. The BERT-based approaches could only give under 50% F1-score on this set, indicating its difficulty.
- The proposed Chinese RC data could also be a resource for cross-lingual research purpose when studied along with SQuAD and other similar datasets.

2 The Proposed Dataset

2.1 Task Definition

Generally, the reading comprehension task can be described as a triple $\langle \mathcal{P}, \mathcal{Q}, \mathcal{A} \rangle$, where $\mathcal{P}$ represents **P**assage, $\mathcal{Q}$ represents **Q**uestion and the $\mathcal{A}$ represents **A**nswer. Specifically, for span-extraction reading comprehension task, the question is annotated by the human, which is much more natural than the cloze-style MRC datasets (Hill et al., 2015; Cui et al., 2016). The answer $\mathcal{A}$ should be a span which is directly extracted from the passage $\mathcal{P}$. According to most of the works on SQuAD, the task can be simplified by predicting the start and end pointer in the passage (Wang and Jiang, 2016).

2.2 Data Pre-Processing

We downloaded Chinese portion of Wikipedia webpage dump[2] on Jan 22, 2018 and used open-source toolkit *Wikipedia Extractor*[3] for pre-processing the raw files into plain text. We also convert the Traditional Chinese characters into Simplified Chinese for normalization purpose using *opencc*[4] toolkit.

2.3 Human Annotation

The questions in the proposed dataset are completely annotated by human experts, which is different from previous works that rely on the automatic data generation (Hermann et al., 2015; Hill et al., 2015; Cui et al., 2016). Before annotating, the document is divided into several passages, and each passage is limited to have no more than 500 Chinese words, where the word is counted by using LTP (Che et al., 2010). Then, the annotator was instructed to first evaluate the appropriateness of the passages, because some of the passages are extremely difficult for the public to understand. Following rules are applied when discarding the passages.

[2] https://dumps.wikimedia.org/zhwiki/latest/

[3] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

[4] https://github.com/BYVoid/OpenCC

- Contain over 30% non-Chinese characters.
- Contain many professional words that hard to understand.
- Contains many special characters and symbols.
- The paragraph is written in classical Chinese, which is substantially different from the Chinese language nowadays.

After identifying the passage is appropriate for annotation, the annotator will read the passage and ask the questions based on it and annotated a primary answer. During the question annotation, the following rules are applied.

- No more than five questions for each passage.
- The answer MUST be a span in the passage to meet the task definition.
- Encourage the question diversity, such as who/when/where/why/how, etc.
- Avoid directly using the description in the passage. Use paraphrase or syntax transformation to add difficulties for answering.
- Long answers (say over 30 characters) will be discarded.

For the evaluation sets, i.e., development, test, challenge, there are three answers available for better evaluation. Besides the primary answer that was annotated by the question proposer, we also invite two additional annotators to write the second and third answers for the question. During this phase, the annotators could not see the primary answer to ensure the answer was not copied from others and encourage the diversities in the answer.

2.4 Challenge Set

In order to examine how well can reading comprehension models deal with the questions that need comprehensive reasoning over various clues in the context, we additionally annotated a small challenge set for this purpose while keeping the span-extraction style. The annotation was also done by three annotators in a similar way that for development and test set. Figure 1 shows an example in the challenge set. The question should meet the following standards to be qualified into this set.

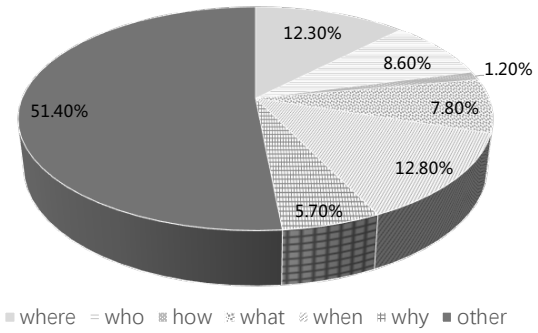

Figure 2: Question types of the development set.

- The answer cannot be only inferred by a single sentence in the passage if the answer is a single word or short phrase. We encourage the annotator to ask the questions that need comprehensive reasoning in the passage to increase the difficulties.
- If the answer belongs to a type of named entity, or specific genre (such as date, color, etc.), it can not be the only one in the context, or the machine could easily pick it out according to its type. For example, if there is only one person name appears in the context, then it cannot be used for annotating questions. There should be at least two person names that could mislead the machine for answering.

2.5 Statistics

The general statistics of the pre-processed data are given in Table 1. The question type distribution of the development set is given in Figure 2.

	Train	Dev	Test	Challenge
Question #	10,321	3,351	4,895	504
Answer per Q	1	**3**	**3**	**3**
Max P tokens	962	961	980	916
Max Q tokens	89	56	50	47
Max A tokens	100	85	92	77
Avg P tokens	452	469	472	464
Avg Q tokens	15	15	15	18
Avg A tokens	17	9	9	19

Table 1: Statistics of the CMRC 2018 dataset. (P: Passage, Q: Question, A: Answer)

3 Evaluation Metrics

In this paper, we adopt two evaluation metrics following Rajpurkar et al. (2016). However, as

	Development		Test		Challenge	
	EM	F1	EM	F1	EM	F1
Estimated Human Performance	*91.083*	*97.348*	*92.400*	*97.914*	*90.382*	*95.248*
Z-Reader (single model)	**79.776**	**92.696**	**74.178**	**88.145**	13.889	**37.422**
MCA-Reader (ensemble)	66.698	85.538	71.175	88.090	15.476	37.104
RCEN (ensemble)	76.328	91.370	68.662	85.753	15.278	34.479
MCA-Reader (single model)	63.902	82.618	68.335	85.707	13.690	33.964
OmegaOne (ensemble)	66.977	84.955	66.272	82.788	12.103	30.859
RCEN (single model)	73.253	89.750	64.576	83.136	10.516	30.994
GM-Reader (ensemble)	58.931	80.069	64.045	83.046	**15.675**	37.315
OmegaOne (single model)	64.430	82.699	64.188	81.539	10.119	29.716
GM-Reader (single model)	56.322	77.412	60.470	80.035	13.690	33.990
R-NET (single model)	45.418	69.825	50.112	73.353	9.921	29.324
SXU-Reader (ensemble)	40.292	66.451	46.210	70.482	N/A	N/A
SXU-Reader (single model)	37.310	66.121	44.270	70.673	6.548	28.116
T-Reader (single model)	39.422	62.414	44.883	66.859	7.341	22.317
BERT-base (Chinese)	63.6	83.9	67.8	86.0	18.4	42.1
BERT-base (Multi-lingual)	64.1	84.4	68.6	86.8	18.6	43.8

Table 2: Baseline results and CMRC 2018 participants' results. Note that, some of the submissions are using development set for training as well.

the Chinese language is quite different from English, we adapt the original metrics in the following ways. Note that, the common punctuations, white spaces are ignored for normalization.

3.1 Exact Match

Measure the exact match between the prediction and ground truths that is 1 for the exact match. Otherwise, the score is 0. This is the same as the one proposed by Rajpurkar et al. (2016).

3.2 F1-Score

Measure the character-level fuzzy match between the prediction and ground truths. Instead of treating the predictions and ground truths as bag-of-words, we calculate the length of the longest common sequence (LCS) between them and compute the F1-score accordingly. We take the maximum F1 over all of the ground truth answers for a given question. Note that, non-Chinese words will not be segmented into characters.

3.3 Estimated Human Performance

We also report the estimated human performance in order to measure the difficulty of the proposed dataset. As we have illustrated in the previous section, there are three answers for each question in development, test, and challenge set. Unlike Rajpurkar et al. (2016), we use a cross-validation method to calculate the performance. We regard the first answer as human prediction and treat the rest of the answers as ground truths. In this way, we can get three human prediction performance by iteratively regarding the first, second, and third answer as the human prediction. Finally, we calculate the average of three results as the final estimated human performance on this dataset.

4 Experimental Results

4.1 Baseline System

Following Devlin et al. (2019), we adopt BERT for our baseline system. Specifically, we slightly modify the `run_squad.py` script[5] for adjusting our dataset, while keeping the most of the original implementation. For the baseline system, we used an initial learning rate of 3e-5 with a batch size of 32 and trained for two epochs. The maximum lengths of document and query are set to 512 and 64.

4.2 Results

The results are shown in Table 2. Besides the baseline systems, we also include the participants' results of CMRC 2018 evaluation. We release the training and development set to the public and accepted submissions from participants to evaluate their models on the hidden test and challenge set to preserve the integrity of the evaluation process following Rajpurkar et al. (2016). As we can see that most of the participants could obtain over 80 in the test F1. While compared to F1 metric, the EM metric is substantially lower compared

[5] https://github.com/google-research/bert/blob/master/run_squad.py

to the SQuAD dataset (usually within 10 points). This suggests that how to determine the exact span boundary in Chinese machine reading comprehension plays a key role to improve the system performance.

Not surprisingly, as shown in the last column of Table 2, though the top-ranked systems obtain decent scores on the development and test set, they are failed to give satisfactory results on the challenge set. However, as we can see that the estimated human performance on the development, test, and challenge set are relatively similar, where the challenge set gives slightly lower scores. We also observed that though Z-Reader obtains best scores on the test set, it failed to give consistent performances on the EM metric of the challenge set. This suggests that the current reading comprehension models are relatively not capable of handling difficult questions that need comprehensive reasoning among several clues in the passage.

BERT-based approaches show competitive performance against participants submissions.[6] Though traditional models have higher scores in the test set, when it comes to the challenge set, BERT-based baselines are consistently higher, demonstrating that rich representation provided by BERT is beneficial for solving harder questions and generalize well among both easy and hard questions.

5 Conclusion

In this work, we propose a span-extraction dataset for Chinese machine reading comprehension. The dataset is annotated by human experts with near 20,000 questions as well as a challenging set which is composed of the questions that need reasoning over multiple clues. The evaluation results show that the machine could give excellent scores on the development and test set with only near 10 points below the estimated human performance in F1-score. However, when it comes to the challenge set, the scores are declining drastically while the human performance remains almost the same with the non-challenge set, indicating that there are still potential challenges in designing more sophisticated models to improve the performance. We hope the release of this dataset could bring language diversity in machine reading comprehension task, and accelerate further investigation on solving the questions that need comprehensive reasoning over multiple clues.

[6] As CMRC 2018 workshop was held before the publication of BERT, systems of the participants are not based on BERT.

Open Challenge

We would like to invite more researchers doing experiments on our CMRC 2018 dataset and evaluate on the hidden test and challenge set to further test the generalization of the models. You can follow the instructions on our CodaLab Worksheet to submit your model via https://bit.ly/2ZdS8Ct

Acknowledgments

We would like to thank the anonymous reviewers for their thorough reviewing and providing thoughtful comments to improve our paper. We would like to thank our resource team for annotating and verifying evaluation data. Also, we thank the Seventeenth China National Conference on Computational Linguistics (CCL 2018)[7] and Changsha University of Science and Technology for providing venue for evaluation workshop. This work was supported by the National Natural Science Foundation of China (NSFC) via grant 61976072, 61632011 and 61772153.

References

Wanxiang Che, Zhenghua Li, and Ting Liu. 2010. Ltp: A chinese language technology platform. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 13–16. Association for Computational Linguistics.

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2017. Attention-over-attention neural networks for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 593–602. Association for Computational Linguistics.

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)*. European Language Resources Association (ELRA).

Yiming Cui, Ting Liu, Zhipeng Chen, Shijin Wang, and Guoping Hu. 2016. Consensus attention-based neural networks for chinese reading comprehension.

[7] http://www.cips-cl.org/static/CCL2018/index.html

In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1777–1786. The COLING 2016 Organizing Committee.

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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2017. Gated-attention readers for text comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1832–1846. Association for Computational Linguistics.

Wei He, Kai Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, et al. 2017. Dureader: a chinese machine reading comprehension dataset from real-world applications. *arXiv preprint arXiv:1711.05073*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1684–1692.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

Minghao Hu, Yuxing Peng, Zhen Huang, Xipeng Qiu, Furu Wei, and Ming Zhou. 2018. Reinforced mnemonic reader for machine reading comprehension. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4099–4106. International Joint Conferences on Artificial Intelligence Organization.

Rudolf Kadlec, Martin Schmid, Ondřej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 908–918. Association for Computational Linguistics.

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*, pages 785–794. Association for Computational Linguistics.

Ting Liu, Yiming Cui, Qingyu Yin, Wei-Nan Zhang, Shijin Wang, and Guoping Hu. 2017. Generating and exploiting large-scale pseudo training data for zero pronoun resolution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 102–111. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 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. Association for Computational Linguistics.

Matthew Richardson, Christopher JC 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*, pages 193–203.

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*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

Shuohang Wang and Jing Jiang. 2016. Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv:1608.07905*.

Wei Wang, Ming Yan, and Chen Wu. 2018. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1705–1714. Association for Computational Linguistics.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198. Association for Computational Linguistics.

Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic coattention networks for question answering. *arXiv preprint arXiv:1611.01604*.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution

with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.