Native Chinese Reader
Supplementary Materials

Project Website: https://sites.google.com/view/native-chinese-reader/

A Project Statement

Dataset: All the materials are collected from online education materials or generated by qualified high-school teachers. All the documents are manually verified by the authors so that they all follow the China's Chinese education standard, which does not contain any sensitive or inappropriate content for high school students. Language of the dataset is written by Mandarin Chinese in the simplified script per China's education system regulation, including the classical Chinese literature.

Accessibility: The dataset, related models as well as our human annotations are all kept at Google drive with links at our project website. All of our authors promise to keep the best efforts to keep them accessible. Our project website also has a detailed introduction to the online competition we organized. We include the requirement that all the participating teams should release their model in our competition agreement, so the release of models are permitted by all the participants.

Human Accuracy: Note that each problem requires non-trivial reasoning and reading a long article. To ensure the students treat the problems seriously, we leave this task as a bonus homework in our deep learning course for second-year college students at Tsinghua university. We also make sure that each student will be assigned no more than 10 articles for a reasonable workload. We also make sure each problem is at least answered by 3 students to get an accurate estimate.

Licence: This dataset is released under the CC BY-SA 4.0 license for general research purposes.

35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.

B Model Details

Our our code can be found at our project website and are released under the MIT license.

The baseline model of this article is based on pre-trained BERT, with some modifications made to the input layer and output layer.

The input of this task consists of three parts, which are articles, questions, and options. There are 2-4 options for each question, we fill the questions into 4 options. For convenience, we denote the document as D, the question as Q, and the answer options as A_1, A_2, A_3, A_4. For the *i*-th option, we construct an input sequence as [CLS] D [SEP] Q [SEP] A_i [SEP]. We truncated the documents by dropping the end part to ensure that the input sequences are not longer than the maximum positions of the pre-trained models.

We fine-tuned several commonly used Chinese pre-trained models as baselines, including Google's BERT-chinese, Baidu's ERNIE, HFL's BERT-wwm, BERT-wwm-ext, and MacBERT [3, 2, 1]. These 5 base models have a similar model structure with 12-layer of Transformer Encoder [4] and 12 attention heads, the hidden size is 768. We also tried two large models, Roberta-large and MacBERT-large, both of which have 24 layers and 16 attention heads, hidden size is 1024.

We compared the results of the 7 baseline models in section 4.2. The BERT-based model will generate a 768-dimensional hidden state for each token of the input sequence. We take the output vector of [CLS] token and map it to a *one-dimensional* logit through a trainable vector. The logit means to what extent A_i may be the correct option. In the same way, we can get the logit of the other 3 options, and employ Softmax to compute the probability of the 4 options. The training objective is to minimize the four-class cross-entropy. In the inferring stage, we just select the option with the highest probability.

All baseline models are trained on 8 Tesla V100 GPU with 32G memory. We set the same hyper-parameters for all base models and large models, separately. For base models, we set epoch as 10, batch size as 64, learning rate as 5e-6. For large models, we set epoch as 10, batch size as 32, learning rate as 2e-6. Regarding data augmentation, we first use the primary school dataset for 5 epoch and then turn to NCR dataset for 5 epochs. All the codes are implemented based on Hugging Face transformers [5]. It takes about 5 minutes to run one epoch for base model and 20 minutes for large model.

The best competition model applied a similar structure. However, they cut the document into several chunks and utilize an information retrieval tool Okapi BM25 to extract the most relevant chunk according to the question as to the actual passage, which reduce the super long document to a reasonable length. The document segment, question and the options are then input to XLNet-based classifier to predict the correct answer. Compared to our baseline model, their IR tool can effectively keep the most informative segment of the document while reducing to a reasonable sequence length.

C Surface pattern in options

We investigate some special patterns and predict the answer based on these patterns. In Table 1, we focus on several absolute quantifiers including "只能" ("can only"), "必然", "必定", "必须", ("must"), "只可能", ("may only"), "绝对" ("absolute"). For each pattern, We keep the questions where only one option contains this pattern or only one option doesn't contain this pattern and choose the special option as the predicted answer. The results are aggregated from the whole dataset with a total of 20477 questions. "combination" represents a combination pattern that including all the aforementioned patterns. We can observe that there are not many questions meeting the requirements, and the accuracy is indeed higher than random but also still very low.

Table 1: Predict the answer based on special patterns. The results are aggregated from the whole dataset with a total of 20477 questions. "combination" represents a combination pattern that including all the aforementioned patterns.

Pattern	只能	必然	必定	必须	只可能	绝对	combination
# Question	291	308	52	658	5	101	1296
Accuracy	0.3470	0.2825	0.3077	0.2903	0	0.2673	0.2994

D Additional Examples

Limited by space, we show some additional examples in this section. The example in Table 2 is a classical Chinese with a question of sentiment. It described a dialogue between the author and his friend, which reflects the author's inner contradictions and complex feelings. The author quoted predecessors' (曹操 Cao Cao) verses and the allusions to **the Battle of Chibi** (赤壁之战) to express his feelings. The question requires the readers to analyze the document from different perspectives, including the author's sentiment, writing techniques, and rhyme.

In Table 3, we present the document of question in Table 6 of the main paper with its English translation, which is a excerpt from Lu Xun's famous article *My old home* (sometimes it is translated to *Hometown*)

References

[1] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Revisiting pre-trained models for chinese natural language processing. *arXiv preprint arXiv:2004.13922*, 2020.

[2] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*, 2019.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[5] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

苏子愀然，正襟危坐，而问客曰：“何为其然也？”客曰：“月明星稀，乌鹊南飞，此非曹孟德之诗乎？西望夏口，东望武昌，山川相缪，郁乎苍苍，此非孟德之困于周郎者乎？方其破荆州，下江陵，顺流而东也，舳舻千里，旌旗蔽空，酾酒临江，横槊赋诗，固一世之雄也，而今安在哉？况吾与子渔樵于江渚之上，侣鱼虾而友麋鹿，驾一叶之扁舟，举匏樽以相属。寄蜉蝣与天地，渺沧海之一粟。哀吾生之须臾，羡长江之无穷。挟飞仙以遨游，抱明月而长终。知不可乎骤得，托遗响于悲风。”苏子曰：“客亦知夫水与月乎？逝者如斯，而未尝往也；盈虚者如彼，而卒莫消长也。盖将自其变者而观之，而天地曾不能一瞬；自其不变者而观之，则物与我皆无尽也，而又何羡乎？且夫天地之间，物各有主，苟非吾之所有，虽一毫而莫取。惟江上之清风，与山间之明月，耳得之而为声，目遇之而成色。取之无禁，用之不竭，是造物者之无尽藏也，而吾与子之所共适。	Sad at heart, I sat up straight to ask my friend why the music was so mournful. He replied, “Didn’t Cao Cao describe a scene like this in his poem: ’The moon is bright, the stars are scattered, the crows fly south...?’ And isn’t this the place where he was defeated by Zhou Yu? See how the mountains and streams intertwine, and how darkly imposing they are with Xiakou to the west and Wuchang to the east. When Cao Cao took Jingzhou by storm and conquered Jiangling, then advanced eastward along the river, his battleships stretched for a thousand li, his armies’ pennons and banners filled the sky. When he offered a libation of wine on the river and lance in hand chanted his poem, he was the hero of his times. But where is he now? We are mere fishermen and woodcutters, keeping company with fish and prawnsand befriending deer. We sail our skiff, frail as a leaf, and toast each other by drinking wine from a gourd. We are nothing but insects who live in this world but one day, mere specks of grain in the vastness of the ocean. I am grieved because our life is so transient, and envy the mighty river which flows on forever. I long to clasp winged fairies and roam freely, or to embrace the bright moon for all eternity. But knowing that this cannot be attained at once, I give vent to my feelings in these notes which pass with the sad breeze. ” Then I asked him, “Have you considered the water and the moon? Water flows away but is never lost; the moon waxes and wanes, but neither increases nor diminished. If you look at its changing aspect, the universe passes in the twinkling of an eye; but if you look at its changeless aspect, all creatures including ourselves are imperishable. What reason have you to envy other things? Besides, everything in this universe has its owner; and if it does not belong to me not a tiny speck can I take. The sole exceptions are the cool breeze on the river, the bright moon over the hills. These serve as music to our ears, as colour to our eyes; these we can take freely and enjoy forever; these are inexhaustible treasures supplied by the Creator, and things in which we can delight together.
Q 下列对文段的理解和分析不正确的一项是	**Q** Choose one from the following options which is incorrect understanding and analysis of the document.
A. 主客对话的内容实际上是作者内心思想矛盾和感情复杂的反映，运用主客问答的方式，可以使行文波澜起伏，摇曳多姿，作者的思想感情也因此充分展现。	A. The content of the dialogue between the author and his friend is actually a reflection of the author’s inner contradictions and complex feelings. The use of question-and-answer methods can make the writing ups and downs, swaying, and fully demonstrate the author’s thoughts and feelings.
B. 作者通过多组对比，揭示出“悲”的原因。文中既写了曹孟德一世之雄的兴亡之悲，也写了由宇宙无穷与人生短暂的对比所生之悲，还写了现实与理想的矛盾所生之悲。	B. The author reveals the cause of "sorrow" through multiple groups of comparisons. The article not only writes the tragedy of the rise and fall of Cao Cao, but also the tragedy born from the contrast between the infinity of the universe and the short-lived life, and the tragedy born from the contradiction between reality and ideal.
C. 作者以江水明月作比，说明世间万物和人生既有变的一面，也有不变的一面，阐述了自然万物变化与永恒的哲理，表现出作者无奈消极的人生态度。⋆	**C. The author uses the river, water and the moon as a comparison to illustrate that everything in the world and life has both a changeable side and an unchanging side. He expounds the change and eternal philosophy of all things in nature, showing the author’s helpless and passive attitude towards life.** ⋆
D. 文段句式骈散结合，结构、句法、韵律都相对自由。大量对偶句的使用使文章参差错落，整齐简约，极富声韵之美。	D. The paragraphs and sentences are combined with parallel and prose, and the structure, syntax, and rhythm are relatively free. The use of a large number of antithetical sentences makes the article jumbled, neat and simple, and full of beauty of rhyme and rhyme.

Table 2: An example of document in Classical Chinese with questions (left) and English translation (right), ⋆ is the correct option. And this question is an example of sentiment

故乡（节选）我这时很兴奋，但不知道怎么说才好，只是说：“阿！闰土哥，——你来了？......”我接着便有许多话，想要连珠一般涌出：角鸡，跳鱼儿，贝壳，猹......但又总觉得被什么挡着似的，单在脑里面回旋，吐不出口外去。他站住了，脸上现出欢喜和凄凉的神情；动着嘴唇，却没有作声。他的态度终于恭敬起来了，分明的叫道：“老爷！......”我似乎打了一个寒噤；我就知道，我们之间已经隔了一层可悲的厚障壁了。我也说不出话。他回过头去说，“水生，给老爷磕头。”便拖出躲在背后的孩子来，这正是一个廿年前的闰土，只是黄瘦些，颈子上没有银圈罢了。“这是第五个孩子，没有见过世面，躲躲闪闪......”母亲和宏儿下楼来了，他们大约也听到了声音。“老太太。信是早收到了。我实在喜欢的了不得，知道老爷回来......”闰土说。“阿，你怎的这样客气起来。你们先前不是哥弟称呼么？还是照旧：迅哥儿。”母亲高兴的说。“阿呀，老太太真是......这成什么规矩。那时是孩子，不懂事......”闰土说着，又叫水生上来打拱，那孩子却害羞，紧紧的只贴在他背后。“他就是水生？第五个？都是生人，怕生也难怪的；还是宏儿和他去走走。”母亲说。宏儿听得这话，便来招水生，水生却松松爽爽同他一路出去了。母亲叫闰土坐，他迟疑了一回，终于就了坐，将长烟管靠在桌旁。

Translation:

My old home (excerpt) **Lu xun** Delighted as I was, I did not know how to express myself, and could only say: "Oh! Jun-tu—so it's you? . . ." After this there were so many things I wanted to talk about, they should have poured out like a string of beads: woodcocks, jumping fish, shells, zha. . . . But I was tongue-tied, unable to put all I was thinking into words. He stood there, mixed joy and sadness showing on his face. His lips moved, but not a sound did he utter. Finally, assuming a respectful attitude, he said clearly: "Master! . . ." I felt a shiver run through me; for I knew then what a lamentably thick wall had grown up between us. Yet I could not say anything. He turned his head to call: "Shui-sheng, bow to the master." Then he pulled forward a boy who had been hiding behind his back, and this was just the Jun-tu of twenty years before, only a little paler and thinner, and he had no silver necklet. "This is my fifth," he said. "He's not used to company, so he's shy and awkward." Mother came downstairs with Hung-erh, probably after hearing our voices. "I got your letter some time ago, madam," said Jun-tu. "I was really so pleased to know the master was coming back. . . ." "Now, why are you so polite? Weren't you playmates together in the past?" said mother gaily. "You had better still call him Brother Hsun as before." "Oh, you are really too. . . . What bad manners that would be. I was a child then and didn't understand." As he was speaking Jun-tu motioned Shui-sheng to come and bow, but the child was shy, and stood stock-still behind his father. "So he is Shui-sheng? Your fifth?" asked mother. "We are all strangers, you can't blame him for feeling shy. Hung-erh had better take him Out to play."

When Hung-eth heard this he went over to Shui-sheng, and Shui-sheng went out with him, entirely at his ease. Mother asked Jun-tu to sir down, and after a little hesitation he did so; leaning his long pipe against the table.

Table 3: The document (top) of example in Table 6 and its English translation. (bottom)

E Datasheet

E.1 Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

we seek to build a native-level Chinese comprehension system, we collect documents with questions from the exam questions for the Chinese course in China's high schools, which are designed to evaluate the language proficiency of native Chinese youth. These questions are also not easy for native Chinese speakers and aim to push the frontier of building native-level Chinese MRC models.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

It was created by Haihua Institute for Frontier Information Technology.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Haihua Institute for Frontier Information Technology.

E.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Each instance contains a document and a list of multiple-choice questions related to this document. And each question is comprised of a question text and 2 4 options, of which exactly one is correct, and the correct answer is also included.

Table 4: Fields and description in NCR

Field	Description
ID	Document ID
Content	Document text
Questions	A list of quetions
Questions:Question	Question text
Qusetions:Answer	The ground truth answer
Questions:Q_id	Question id
Type (Annotated in validation/test set)	Writing style of document, 00 for moder-style (without poetry), 11 for classical-style (without poetry), 22 for classical poetry, 33 for modern poetry

How many instances are there in total (of each type, if appropriate)?

NCR consists of 6315 documents with 15419 questions for training, 1000 documents with 2443 questions for validation and 1073 documents with 2615 questions for testing.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is a sample of instances from the larger set containing all exam questions for the Chinese course in China's high schools. It is impossible to collect all the questions since there are too many questions online and many new questions will be generated every day.

What data does each instance consist of? “Raw” data (e.g., unprocessed text or images)or features? In either case, please provide a description.

Data is all in the form of text.

Is there a label or target associated with each instance? If so, please provide a description.

We provide each question with the correct answer. For data in the validation and test set, we also provide the writing style of the document.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

There is no information missing.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

A document is associated with several questions, these questions share the same document.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

We randomly split the dataset collected online at the document level, with 6315 for training, 1000 for validation and 1000 for testing. To make sure our test set has sufficient novel questions that never appear online, we also invited a few high-school Chinese teachers to manually generate 193 questions for a total of 73 additional documents to augment the test set. Finally NCR consists of 6315 documents with 15419 questions for training, 1000 documents with 2443 questions for validation and 1073 documents with 2615 questions for testing.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

There are no errors, sources of noise, or redundancies in the dataset because we filter out these data.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained.

Does the dataset contain data that might be considered confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

No, all data is public.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

The dataset does not contain any data that may be offensive, insulting, threatening, or might otherwise cause anxiety.

Does the dataset relate to people? If not, you may skip the remaining questions in this subsection.

This dataset does not relate to people.

E.3 Collection

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

A document is associated with several questions, these questions share the same document, which is directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data was collected by manual human curation. And the collected data was validated by another person to ensure the quality.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

We contracted out the data collection process to 海天瑞声 (SpeechOcean, http://en.speechocean.com/)

Over what timeframe was the data collected? Does this timeframe match the creation time frame of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data collection process lasted for around 40 days. Since the dataset contains many articles in classical Chinese, which can be date back to thousands of years ago.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No ethical review processes were conducted.

E.4 Processing

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this subsection.

We clean the data by filter out some questions which are base on the format. For example, some questions are about the marked or boldface words.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

No.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

No.

E.5 Uses

Has the dataset been used for any tasks already? If so, please provide a description.

To examine the limit of current MRC methods, we organized a 3-month-long online competition using NCR with a training and validation set released. Participants are allowed to use any open-access pre-trained model or any open-access unlabeled data. Use of any external MRC supervision is forbidden, since a portion of the test questions are possibly accessible online. This aims to prevent human annotations overlapping with our held-out data for a fair competition. There are a total of 141 participating teams and the best submission model with the highest test accuracy is taken as the competition model. The team is from an industry lab. They first pre-trained an XLNet-based model on a company-collected large corpus. For each question, they use an information retrieval tool Okapi BM25 to extract the most relevant parts from the document and then run this pre-trained model for answer selection based on the extracted texts. The final model with the highest accuracy is released: `https://github.com/xssstory/NCR_competition_model`

Is there a repository that links to any or all papers or systems that use the dataset?

The competition website: `https://www.biendata.xyz/competition/haihua_2021/`

Github for the competition model with the highest accuracy: `https://github.com/xssstory/NCR_competition_model`

Github for baselines: `https://github.com/xssstory/NCR_baseline`

What (other) tasks could the dataset be used for?

This dataset may also be used for a question-answering generation task.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

Users should keep in mind that the questions and their answer comes from different teachers.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Unknown.

E.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

Yes, the dataset is public now.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset is available at `https://github.com/xssstory/NCR_competition_model` and `https://drive.google.com/drive/folders/1Ci-KLHKk-yP-y5fWX4_cU8bA2fL_q76e?usp=sharing`

When will the dataset be distributed?

It is available now.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

This dataset is released under the CC BY-SA 4.0 license for general research purposes.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No

E.7 Maintance

Who is supporting/hosting/maintaining the dataset?

The dataset will be hosted on GitHub and google drive and will be maintained by Shusheng Xu and Yichen Liu.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The maintainers can be contacted at xuss20@mails.tsinghua.edu.cn and yl7043@nyu.edu

Is there an erratum? If so, please provide a link or other access point.

No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

The dataset will be updated if necessary, updates will be communicated via the project website `https://sites.google.com/view/native-chinese-reader/` Github at `https://github.com/xssstory/NCR_competition_model` and `https://github.com/xssstory/NCR_baseline`

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its will be communicated to users.

Yes, it will be communicated to users via the project website `https://sites.google.com/view/native-chinese-reader/` Github at `https://github.com/xssstory/NCR_competition_model` and `https://github.com/xssstory/NCR_baseline`

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

This is not supported now.