Dataset for the First Evaluation on Chinese Machine Reading Comprehension

Yiming Cui[†], Ting Liu[‡], Zhipeng Chen[†], Wentao Ma[†], Shijin Wang[†] and Guoping Hu[†]
[†]Joint Laboratory of HIT and iFLYTEK, iFLYTEK Research, Beijing, China
[‡]Research Center for Social Computing and Information Retrieval,
Harbin Institute of Technology, Harbin, China
[†]{ymcui,zpchen,wtma,sjwang3,gphu}@iflytek.com
[‡]tliu@ir.hit.edu.cn

Abstract

Machine Reading Comprehension (MRC) has become enormously popular recently and has attracted a lot of attention. However, existing reading comprehension datasets are mostly in English. To add diversity in reading comprehension datasets, in this paper we propose a new Chinese reading comprehension dataset for accelerating related research in the community. The proposed dataset contains two different types: cloze-style reading comprehension and user query reading comprehension, associated with large-scale training data as well as human-annotated validation and hidden test set. Along with this dataset, we also hosted the first Evaluation on Chinese Machine Reading Comprehension (CMRC-2017) and successfully attracted tens of participants, which suggest the potential impact of this dataset.

Keywords: Chinese Reading Comprehension, Question Answering, Evaluation

1. Introduction

Machine Reading Comprehension (MRC) has become enormously popular in recent research, which aims to teach the machine to comprehend human languages and answer the questions based on the reading materials. Among various reading comprehension tasks, the cloze-style reaing comprehension is relatively easy to follow due to its simplicity in definition, which requires the model to fill an exact word into the query to form a coherent sentence according to the document material. Several cloze-style reading comprehension datasets are publicly available, such as CNN/Daily Mail (Hermann et al., 2015), Children's Book Test (Hill et al., 2015), People Daily and Children's Fairy Tale (Cui et al., 2016).

In this paper, we provide a new Chinese reading comprehension dataset[1], which has the following features

- We provide a large-scale automatically generated Chinese cloze-style reading comprehension dataset, which is gathered from children's reading material.
- Despite the automatic generation of training data, our evaluation datasets (validation and test) are annotated manually, which is different from previous works.
- To add more diversity and for further investigation on transfer learning, we also provide another evaluation datasets which is also annotated by human, but the query is more natural than the cloze type.

We also host the 1st Evaluation on Chinese Machine Reading Comprehension (CMRC2017), which has attracted over 30 participants and finally there were 17 participants submitted their evaluation systems for testing their reading comprehension models on our newly developed dataset, suggesting its potential impact. We hope the release of the dataset to the public will accelerate the progress of Chinese research community on machine reading comprehension field.

We also provide four official baselines for the evaluations, including two traditional baselines and two neural baselines. In this paper, we adopt two widely used neural reading comprehension model: AS Reader (Kadlec et al., 2016) and AoA Reader (Cui et al., 2017).

The rest of the paper will be organized as follows. In Section 2, we will introduce the related works on the reading comprehension dataset, and then the proposed dataset as well as our competitions will be illustrated in Section 3. The baseline and participant system results will be given in Section 4 and we will made a brief conclusion at the end of this paper.

2. Related Works

In this section, we will introduce several public cloze-style reading comprehension dataset.

2.1. CNN/Daily Mail

Some news articles often come along with a short summary or brief introduction. Inspired by this, Hermann et al. (2015) release the first cloze-style reading comprehension dataset, called CNN/Daily Mail[2]. Firstly, they obtained large-scale CNN and Daily Mail news data from online websites, including main body and its summary. Then they regard the main body of the news as the *Document*. The *Query* is generated by replacing a name entity word from the summary by a placeholder, and the replaced named entity word becomes the *Answer*. Along with the techniques illustrated above, after the initial data generation, they also propose to anonymize all named entity tokens in the data to avoid the model exploit world knowledge of specific entities, increasing the difficulties in this dataset. However,

[1] CMRC 2017 Public Datasets: https://github.com/ymcui/cmrc2017.

[2] The pre-processed CNN and Daily Mail datasets are available at http://cs.nyu.edu/~kcho/DMQA/

		Cloze Track		User Query Track	
	Train	Validation	Test	Validation	Test
# Query	354,295	2,000	3,000	2,000	3,000
Max # tokens in docs	486	481	484	481	486
Max # tokens in query	184	72	106	21	29
Avg # tokens in docs	324	321	307	310	290
Avg # tokens in query	27	19	23	8	8
Vocabulary			94,352		

Table 1: Statistics of the dataset for the 1st Evaluation on Chinese Machine Reading Comprehension (CMRC-2017).

as we have known that world knowledge is very important when we do reading comprehension in reality, which makes this dataset much artificial than real situation. Chen et al. (2016) also showed that the proposed anonymization in CNN/Daily Mail dataset is less useful, and the current models (Kadlec et al., 2016; Chen et al., 2016) are nearly reaching ceiling performance with the automatically generated dataset which contains much errors, such as coreference errors, ambiguous questions etc.

2.2. Children's Book Test

Another popular cloze-style reading comprehension dataset is the Children's Book Test (CBT)[3] proposed by Hill et al. (2015) which was built from the children's book stories. Though the CBT dataset also use an automatic way for data generation, there are several differences to the CNN/Daily Mail dataset. They regard the first 20 consecutive sentences in a story as the *Document* and the following 21st sentence as the *Query* where one token is replaced by a placeholder to indicate the blank to fill in. Unlike the CNN/Daily Mail dataset, in CBT, the replaced word are chosen from various types: Name Entity (NE), Common Nouns (CN), Verbs (V) and Prepositions (P). The experimental results showed that, the verb and preposition answers are not sensitive to the changes of document, so the following works are mainly focusing on solving the NE and CN genres.

2.3. People Daily & Children's Fairy Tale

The previously mentioned datasets are all in English. To add diversities to the reading comprehension datasets, Cui et al. (2016) proposed the first Chinese cloze-style reading comprehension dataset: People Daily & Children's Fairy Tale, including People Daily news datasets and Children's Fairy Tale datasets. They also generate the data in an automatic manner, which is similar to the previous datasets. They choose short articles (several hundreds of words) as *Document* and remove a word from it, whose type is mostly named entities and common nouns. Then the sentence that contains the removed word will be regarded as *Query*. To add difficulties to the dataset, along with the automatically generated evaluation sets (validation/test), they also release a human-annotated evaluation set. The experimental results show that the human-annotated evaluation set is significantly harder than the automatically generated questions. The reason would be that the automatically generated data is accordance with the training data which is also automatically generated and they share many similar characteristics, which is not the case when it comes to human-annotated data.

3. The Proposed Dataset

In this section, we will briefly introduce the evaluation tracks and then the generation method of our dataset will be illustrated in detail.

3.1. The 1st Evaluation on Chinese Machine Reading Comprehension (CMRC-2017)

The proposed dataset is typically used for the 1st Evaluation on Chinese Machine Reading Comprehension (CMRC-2017)[4], which aims to provide a communication platform to the Chinese communities in the related fields. In this evaluation, we provide two tracks. We provide a shared training data for both tracks and separated evaluation data.

- Cloze Track: In this track, the participants are required to use the large-scale training data to train their cloze system and evaluate on the cloze evaluation track, where training and test set are exactly the same type.
- User Query Track: This track is designed for using transfer learning or domain adaptation to minimize the gap between cloze training data and user query evaluation data, i.e. training and testing is fairly different.

Following Rajpurkar et al. (2016), we preserve the test set only visible to ourselves and require the participants submit their system in order to provide a fair comparison among participants and avoid tuning performance on the test set. The examples of Cloze and User Query Track are given in Figure 1.

3.2. Definition of Cloze Task

The cloze-style reading comprehension can be described as a triple $\langle \mathcal{D}, \mathcal{Q}, \mathcal{A} \rangle$, where $\mathcal{D}$ represents **D**ocument, $\mathcal{Q}$ represents **Q**uery and the $\mathcal{A}$ represents **A**nswer. There is a restriction that the answer should be a single word and should appear in the document, which was also adopted in (Hill et al., 2015; Cui et al., 2016). In our dataset, we mainly focus on answering common nouns and named entities which require further comprehension of the document.

[3] Available at `http://www.thespermwhale.com/jaseweston/babi/CBTest.tgz`

[4] CMRC 2017 Official Website: `http://www.hfl-tek.com/cmrc2017/index.html`.

<table>
<tr><th></th><th>Cloze Track</th><th>User Query Track</th></tr>
<tr><td><b>Document</b></td><td>1 ||| 为了 让 森林 变 得 更加 茂盛 ，大伙 都 在 努力 地 工作 着 。
1 ||| To let the forest become more lush, every is working hard.
2 ||| <b>XXXXX</b> 每天 天 不 亮 就 起来 ，在 这 棵 树 上 啄啄 ，在 那 棵 树上 敲敲 ，他 的 尖利 的 长嘴 ，使 害虫 没有 藏身 之 地 。
2 ||| <b>XXXXX</b> wake up very early everyday, digging the tree trunk with its sharp beak and there is no hiding place for the insects.
……
5 ||| 惟有 大象 没 活 干 ，他 整天 游荡 ，大家 问 他 为什么 不 干活 ，他 说 ：“我 没有 <b><u>啄木鸟</u></b> 的 长嘴 ，也 没有 猴子 的 巧手 和 松鼠 的 尖牙利爪 ，我 能 干 什么 呢？
5 ||| He said: “I have no beak as <b><u>woodpecker</u></b> and no hands like monkey or squirrel. So what can I do?”
…
13 ||| 不久 ，在 原先 堆 着 枯树干 的 地方 ，长 出 一 支 支 小 绿苗 。
13 ||| Soon, there are a few green seedlings where the dead tree trunk piled.
14 ||| 大伙 夸奖 大象 有 一 只 多么 能干 的 鼻子 。
14 ||| Everyone praised the elephant that he has a competent nose.</td><td>1 ||| 一 只 <b><u>驴子</u></b> ，捎 着 木料 ，向前 走 去 。
1 ||| The <b><u>donkey</u></b> is going forward carrying wood.
2 ||| 一不小心 ，摔 在 池 里 ，辗转 不 能 出 水 ，他 便 唉声叹气 地 悲哀 起 来 。
2 ||| Accidentally, it fell into the pool and cannot get out from it with sad sighs.
3 ||| 许多 蛙 ，是 生惯 在 池里 的 ，他们 听见 了 驴子 的 呼救声 ，都 来 围 观 。
3 ||| Many frogs are used to live in the pool. They heard the cry for help and head for that.
4 ||| 他们 对 驴子 说 ：" 你 不过 在 池 里 只 一刻儿 功夫 ，便 这样 地 大 嚷 着 救命 ，请 告诉 我们 这 是 什么 缘故 ？
4 ||| They talk to the donkey: “Why you are shouting for staying in the water only for a moment?”
5 ||| 万一 你 像 我们 一样 无穷期 地 居住 在 这里 ，你 又 得 怎样 呢 ？
5 ||| What if you live here like us?
6 ||| " 这 便是 群蛙 给予 驴子 的 讥刺 的 慰藉 。
6 ||| The frogs are giving gibing comfort to the donkey.</td></tr>
<tr><td><b>Query</b></td><td><b>XXXXX</b> 每天 天 不 亮 就 起来 ，在 这 棵 树 上 啄啄 ，在 那 棵 树上 敲敲 ，他 的 尖利 的 长嘴 ，使 害虫 没有 藏身 之 地 。
<b>XXXXX</b> wake up very early everyday, digging the tree trunk with its sharp beak and there is no hiding place for the insects.</td><td>谁 在 大 嚷 着 救命 ？
Who was shouting for help?</td></tr>
<tr><td><b>Answer</b></td><td>啄木鸟
Woodpecker</td><td>驴子
Donkey</td></tr>
</table>

Figure 1: Examples of the proposed datasets (the English translation is in grey). The sentence ID is depicted at the beginning of each row. In the Cloze Track, “XXXXX” represents the missing word.

3.3. Automatic Generation

Following Cui et al. (Cui et al., 2016), we also use similar way to generate our training data automatically. Firstly we roughly collected 20,000 passages from children's reading materials which were crawled in-house. Briefly, we choose an answer word in the document and treat the sentence containing answer word as the query, where the answer is replaced by a placeholder “XXXXX”. The detailed procedures can be illustrated as follows.

- **Pre-processing**: For each sentence in the document, we do word segmentation, POS tagging and dependency parsing using LTP toolkit (Che et al., 2010).
- **Dependency Extraction**: Extract following dependencies: COO, SBV, VOB, HED, FOB, IOB, POB[5], and only preserve the parts that have dependencies.
- **Further Filtering**: Only preserve SBV, VOB and restrict the related words not to be pronouns and verbs.
- **Frequency Restriction**: After calculating word frequencies, only word frequency that greater than 2 is valid for generating question.
- **Question Restriction**: Only five questions can be extracted within one passage.

3.4. Human Annotation

Apart from the automatically generated large-scale training data, we also provide human-annotated validation and test data to improve the estimation quality. The annotation procedure costs one month with 5 annotators and each question is cross-validated by another annotator. The detailed procedure for each type of dataset can be illustrated as follows.

3.4.1. Cloze-style Reading Comprehension

For the validation and test set in cloze data, we first randomly choose 5,000 paragraphs each for automatically generating questions using the techniques mentioned above. Then we invite our resource team to manually select 2,000 questions based on the following rules.

- Whether the question is appropriate and correct
- Whether the question is hard for LMs to answer
- Only select one question for each paragraph

3.4.2. User Query Reading Comprehension

Unlike the cloze dataset, we have no automatic question generation procedure in this type. In the user query dataset, we asked our annotator to directly raise questions according to the passage, which is much difficult and time-consuming than just selecting automatically generated questions. We also assign 5,000 paragraphs for question annotations in both validation and test data. Following rules are applied in asking questions.

- The paragraph should be read carefully and judged whether appropriate for asking questions
- No more than 5 questions for each passage
- The answer should be better in the type of nouns, named entities to be fully evaluated
- Too long or too short paragraphs should be skipped

4. Experiments

In this section, we will give several baseline systems for evaluating our datasets as well as presenting several top-ranked systems in the competition.

[5] Full descriptions of abbreviations can be found at `http://www.ltp-cloud.com/intro/en/#dp_how`.

		Single Model		Ensemble	
Rank	System	Validation	Test	Validation	Test
-	Baseline - Random Guess	1.65	1.67	-	-
-	Baseline - Top Frequency	14.85	14.07	-	-
-	Baseline - AS Reader (default settings)	76.05	77.67	-	-
-	Baseline - AoA Reader (default settings)	**77.20**	**78.63**	-	-
1	6ESTATES	75.85	74.73	**81.85**	**81.90**
2	Shanghai Jiao Tong Univeristy BCMI-NLP	76.15	**77.73**	78.35	80.67
3	XinkTech	77.15	77.53	79.20	80.27
4	East China Normal University (ECNU)	**77.95**	77.40	79.45	79.70
5	Ludong University	74.75	75.07	77.05	77.07
6	Wuhan University (WHU)	78.20	76.53	-	-
7	Harbin Institute of Technology at Shenzhen (HITSZ)	76.05	75.93	-	-
8	HuoYan Technology	73.55	75.77	-	-
9	Wuhan University of Science and Technology (WUST)	73.80	74.53	-	-
10	Beijing Information Science and Technology University	70.05	70.20	-	-
11	Shanxi Univerisity (SXU-2)	62.60	64.70	66.65	68.47
12	Shenyang Aerospace University (SAU)	63.15	65.80	-	-
13	Shanxi University (SXU-1)	64.85	64.67	-	-
14	Zhengzhou Univerisity (ZZU)	52.80	54.53	-	-

Table 2: Results on Cloze Track. The best baseline and participant systems are depicted in bold face.

4.1. Baseline Systems

We set several baseline systems for testing basic performance of our datasets and provide meaningful comparisons to the participant systems. In this paper, we provide four baseline systems, including two simple ones and two neural network models. The details of the baseline systems are illustrated as follows.

- **Random Guess**: In this baseline, we randomly choose one word in the document as the answer.
- **Top Frequency**: We choose the most frequent word in the document as the answer.
- **AS Reader**: We implemented Attention Sum Reader (AS Reader) (Kadlec et al., 2016) for modeling document and query and predicting the answer with the Pointer Network (Vinyals et al., 2015), which is a popular framework for cloze-style reading comprehension. Apart from setting embedding and hidden layer size as 256, we did not change other hyper-parameters and experimental setups as used in Kadlec et al. (2016), nor we tuned the system for further improvements.
- **AoA Reader**: We also implemented Attention-over-Attention Reader (AoA Reader) (Cui et al., 2017) which is the state-of-the-art model for cloze-style reading comprehension. We follow hyper-parameter settings in AS Reader baseline without further tuning.

In the User Query Track, as there is a gap between training and validation, we follow (Liu et al., 2017) and regard this task as domain adaptation or transfer learning problem. The neural baselines are built by the following steps.

System	Validation	Test
Baseline - Random Guess	1.50	1.47
Baseline - Top Frequency	10.65	8.73
Baseline - AS Reader	-	49.03
Baseline - AoA Reader	-	**51.53**
ECNU (Ensemble)	90.45	**69.53**
ECNU (single model)	85.55	65.77
Shanxi University (Team-3)	47.80	49.07
Zhengzhou University	31.10	32.53

Table 3: Results on User Query Track. Due to the using of validation data, we did not report its performance.

- We first use the shared training data to build a general systems, and choose the best performing model (in terms of cloze validation set) as baseline.
- Use User Query validation data for further tuning the systems with 10-fold cross-validations.
- Increase dropout rate (Srivastava et al., 2014) to 0.5 for preventing over-fitting issue.

All baseline systems are chosen according to the performance of the validation set.

4.2. Participant Systems

The participant system results[6] are given in Table 2 and 3.

4.2.1. Cloze Track

As we can see that two neural baselines are competitive among participant systems and AoA Reader successfully

[6] Full CMRC 2017 Leaderboard: `http://www.hfl-tek.com/cmrc2017/leaderboard.html`.

outperform AS Reader and all participant systems in single model condition, which proves that it is a strong baseline system even without further fine-tuning procedure. Also, the best performing single model among participant systems failed to win in the ensemble condition, which suggest that choosing right ensemble method is essential in most of the competitions and should be carefully studied for further performance improvements.

4.2.2. User Query Track

Not surprisingly, we only received three participant systems in User Query Track, as it is much difficult than Cloze Track. As shown in Table 3, the test set performance is significantly lower than that of Cloze Track, due to the mismatch between training and test data. The baseline results give competitive performance among three participants, while failed to outperform the best single model by ECNU, which suggest that there is much room for tuning and using more complex methods for domain adaptation.

5. Conclusion

In this paper, we propose a new Chinese reading comprehension dataset for the 1st Evaluation on Chinese Machine Reading Comprehension (CMRC-2017), consisting large-scale automatically generated training set and human-annotated validation and test set. Many participants have verified their algorithms on this dataset and tested on the hidden test set for final evaluation. The experimental results show that the neural baselines are tough to beat and there is still much room for using complicated transfer learning method to better solve the User Query Task. We hope the release of the full dataset (including hidden test set) could help the participants have a better knowledge of their systems and encourage more researchers to do experiments on.

6. Acknowledgements

We would like to thank the anonymous reviewers for their thorough reviewing and providing thoughtful comments to improve our paper. We thank the Sixteenth China National Conference on Computational Linguistics (CCL 2017) and Nanjing Normal University for providing space for evaluation workshop. Also we want to thank our resource team for annotating and verifying evaluation data. This work was supported by the National 863 Leading Technology Research Project via grant 2015AA015409.

7. Bibliographical References

Che, W., Li, Z., and Liu, T. (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.

Chen, D., Bolton, J., and Manning, D. C. (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 (Volume 1: Long Papers)*, pages 2358–2367. Association for Computational Linguistics.

Cui, Y., Liu, T., Chen, Z., Wang, S., and Hu, G. (2016). Consensus attention-based neural networks for chinese reading comprehension. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1777–1786. The COLING 2016 Organizing Committee.

Cui, Y., Chen, Z., Wei, S., Wang, S., Liu, T., and Hu, G. (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.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1684–1692.

Hill, F., Bordes, A., Chopra, S., and Weston, J. (2015). The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

Kadlec, R., Schmid, M., Bajgar, O., and Kleindienst, J. (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.

Liu, T., Cui, Y., Yin, Q., Zhang, W.-N., Wang, S., and Hu, G. (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.

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. Association for Computational Linguistics.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Vinyals, O., Fortunato, M., and Jaitly, N. (2015). Pointer networks. In *Advances in Neural Information Processing Systems*, pages 2692–2700.