A Span-Extraction Dataset for Chinese Machine Reading Comprehension

Yiming Cui†‡, Ting Liu†, Li Xiao†, Zhipeng Chen†, Wentao Ma†, Wanxiang Che†, Shijin Wang†, Guoping Hu†
†Joint Laboratory of HIT and iFLYTEK (HFL), iFLYTEK Research, Beijing, China
‡Research Center for Social Computing and Information Retrieval (SCIR), Harbin Institute of Technology, Harbin, China
†{ymcui,lixiao3,zpchen,wtma,sjwang3,gphu}@iflytek.com
‡{ymcui,tliu,car}@ir.hit.edu.cn

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 by human on Wikipedia paragraphs. We also annotated a challenge set which contains the questions that need comprehensive understanding and multi-sentence inference throughout the context. 

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 machine reading comprehension research in Chinese language.

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 make 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; 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 Daily & Children’s Fairy Tale, including People Daily news datasets and Children’s Fairy Tale datasets. To add more diversity and for further investigation on transfer learning, they also provide another evaluation dataset which is also annotated by human, 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, which contains 200k queries annotated from the user query logs on search engine.

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 similar performance on the dataset of different languages. To further accelerate the development of the machine reading comprehension research, we propose a span-extraction dataset for Chinese machine reading comprehension.

The main contributions of our work can be con-
2.1 Task Definition

The reading comprehension can be described as a triple \((D, Q, A)\), where \(D\) represents Document, \(Q\) represents Question and the \(A\) represents Answer. Specifically, for span-extraction reading comprehension task, the question is annotated by the human which is much more natural than the cloze-style reading comprehension dataset (Hill et al., 2015; Cui et al., 2016). The answer \(A\) should be directly extracted from the document \(D\). According to the most of the works on SQuAD dataset, the task can be simplified by predicting the start and end point in the document to get the answer span (Wang and Jiang, 2016).

2.2 Data Pre-processing

We downloaded Chinese portion of Wikipedia webpage dump on Jan 22, 2018\(^2\) and use open-source toolkit Wikipedia Extractor\(^3\) for preprocessing the 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 totally annotated by human which is different from previous works that relies 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 paragraphs and each paragraph 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 document, because some of the documents are extremely difficult for the public to understand. Following rules are applied when discarding the document.

- The paragraph contains many non-Chinese characters, say over 30
- The paragraph contains many professional words that hard to understand.
- The paragraph contains many special characters (may induced by the pre-processing process).

---

\(^1\)https://dumps.wikimedia.org/zhwiki/latest/
\(^2\)http://medialab.di.unipi.it/wiki/Wikipedia_Extractor
\(^3\)https://github.com/BYVoid/OpenCC
The paragraph is written in classical Chinese.

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

- No more than 5 questions for each paragraph.
- The answer MUST be a span in the paragraph to meet the task definition.
- Encourage the question diversity in various types, such as who/when/where/why/how etc.
- Avoid directly using the description in the paragraph. Use paraphrase or syntax transformation to add difficulties for answering.
- Long answers (> 30 characters) will be discarded.

For the evaluation sets (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 annotator to write the second and third answer for the question. During this phase, the annotator 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. The annotation was also done by three annotators in a similar way that for development and test set annotation. The question should meet the following standard to be qualified into this set.

- The answer cannot be inferred by a single sentence in the context, if the answer is only a single word or short phrase.
- If the answer belongs to a type of named entity, it can not be the only one in the context, or the machine could easily pick it out according to its named entity 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.

The Figure 2 show an example of the challenge dataset.

2.5 Statistics

The general statistics of the pre-processed data is given in Table 1. The query type distribution of the development set is given in Figure 3.

3 Evaluation Metrics

In this paper, we adopt two evaluation metrics following Rajpurkar et al. (2016). However, Chinese language is fairly different from English, we adapt the original metrics in the following ways. Note that, the common punctuations, white spaces are ignored.

3.1 Exact Match

Measure the exact match between the prediction and ground truths, i.e. 1 for exact match, otherwise the score is 0.
<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Challenge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question #</td>
<td>10,321</td>
<td>3,351</td>
<td>4,895</td>
<td>504</td>
</tr>
<tr>
<td>Answer # per query</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Max doc tokens</td>
<td>962</td>
<td>961</td>
<td>980</td>
<td>916</td>
</tr>
<tr>
<td>Max question tokens</td>
<td>89</td>
<td>56</td>
<td>50</td>
<td>47</td>
</tr>
<tr>
<td>Max answer tokens</td>
<td>100</td>
<td>85</td>
<td>92</td>
<td>77</td>
</tr>
<tr>
<td>Average doc tokens</td>
<td>452</td>
<td>469</td>
<td>472</td>
<td>464</td>
</tr>
<tr>
<td>Average question tokens</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td>Average answer tokens</td>
<td>17</td>
<td>9</td>
<td>9</td>
<td>19</td>
</tr>
</tbody>
</table>

Table 1: Statistics of the proposed CMRC 2018 dataset.

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 bags 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.

3.3 Human Performance

We also report the 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 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 human prediction. Finally, we calculate the average of three results as the final human performance on this dataset.

4 Results

4.1 CMRC 2018 Evaluation Results

The final leaderboard of the CMRC 2018 evaluation is shown in Table 2. As we can see that most of the systems could obtain over 80 in test F1. While compared to F1 metric, the EM metric is substantially lower compared to the SQuAD dataset (usually within 10 points). This suggests that how to determine the exact boundary in Chinese machine reading comprehension plays a key role to improve the system performance.

4.2 Results on the Challenge Set

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 the same, 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 performance on the EM metric of the challenge set. This suggests that the current reading comprehension models are not capable of handling difficult questions that need comprehensive reasoning among several clues in the context.

5 Conclusion

In this work, we propose a span-extraction dataset for Chinese machine reading comprehension, namely CMRC 2018, which is also used in the second evaluation workshop on Chinese machine reading comprehension. The dataset is annotated
by human with near 20,000 questions as well as a challenging set which is composed by 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.

**Open Challenge**

We would like to invite more researchers doing experiments on CMRC 2018 datasets 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://worksheets.codalab.org/worksheets/0x92a80d2fab4f7f9a2b4064f7ddca9ce/](https://worksheets.codalab.org/worksheets/0x92a80d2fab4f7f9a2b4064f7ddca9ce/).

**Acknowledgements**

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) and Changsha University of Science and Technology for providing venue for evaluation workshop. This work was supported by the National 863 Leading Technology Research Project via grant 2015AA015409.

**References**


### Table 2: Overall leaderboard on the CMRC 2018 dataset. The ranking is obtained by the average score of the test EM and F1 in descending order. The results marked as ‘N/A’ are due to the bundle deletion by the participants.

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Development EM</th>
<th>Development F1</th>
<th>Test EM</th>
<th>Test F1</th>
<th>Challenge EM</th>
<th>Challenge F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z-Reader (single model)</td>
<td>79.776</td>
<td>92.696</td>
<td>74.178</td>
<td>88.145</td>
<td>13.889</td>
<td>37.422</td>
</tr>
<tr>
<td>MCA-Reader (ensemble)</td>
<td>66.698</td>
<td>85.538</td>
<td>71.175</td>
<td>88.090</td>
<td>15.476</td>
<td>37.104</td>
</tr>
<tr>
<td>RCEN (ensemble)</td>
<td>76.328</td>
<td>91.370</td>
<td>68.662</td>
<td>85.753</td>
<td>15.278</td>
<td>34.479</td>
</tr>
<tr>
<td>MCA-Reader (single model)</td>
<td>63.902</td>
<td>82.618</td>
<td>68.335</td>
<td>85.707</td>
<td>12.013</td>
<td>30.859</td>
</tr>
<tr>
<td>OmegaOne (ensemble)</td>
<td>66.977</td>
<td>84.955</td>
<td>66.272</td>
<td>82.788</td>
<td>10.119</td>
<td>29.716</td>
</tr>
<tr>
<td>RCEN (single model)</td>
<td>73.253</td>
<td>89.750</td>
<td>64.576</td>
<td>83.136</td>
<td>10.516</td>
<td>30.994</td>
</tr>
<tr>
<td>GM-Reader (ensemble)</td>
<td>58.931</td>
<td>80.069</td>
<td>64.045</td>
<td>83.046</td>
<td>15.675</td>
<td>37.315</td>
</tr>
<tr>
<td>OmegaOne (single model)</td>
<td>64.430</td>
<td>82.699</td>
<td>64.188</td>
<td>81.539</td>
<td>10.119</td>
<td>29.716</td>
</tr>
<tr>
<td>GM-Reader (single model)</td>
<td>56.322</td>
<td>77.412</td>
<td>60.470</td>
<td>80.035</td>
<td>13.690</td>
<td>33.990</td>
</tr>
<tr>
<td>R-NET (single model)</td>
<td>45.418</td>
<td>69.825</td>
<td>50.112</td>
<td>73.353</td>
<td>9.921</td>
<td>29.324</td>
</tr>
<tr>
<td>SXU-Reader (ensemble)</td>
<td>40.292</td>
<td>66.451</td>
<td>46.210</td>
<td>70.482</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>SXU-Reader (single model)</td>
<td>37.310</td>
<td>66.121</td>
<td>44.270</td>
<td>70.673</td>
<td>6.548</td>
<td>28.116</td>
</tr>
<tr>
<td>T-Reader (single model)</td>
<td>39.422</td>
<td>62.414</td>
<td>44.883</td>
<td>66.859</td>
<td>7.341</td>
<td>22.317</td>
</tr>
<tr>
<td>Unnamed Sys by usst (single model)</td>
<td>34.490</td>
<td>59.539</td>
<td>37.916</td>
<td>63.502</td>
<td>5.159</td>
<td>18.687</td>
</tr>
<tr>
<td>Unnamed Sys by whu (single model)</td>
<td>18.577</td>
<td>42.560</td>
<td>22.288</td>
<td>46.774</td>
<td>2.183</td>
<td>21.587</td>
</tr>
<tr>
<td>Unnamed Sys by LittleBai (single model)</td>
<td>7.021</td>
<td>31.657</td>
<td>10.848</td>
<td>37.231</td>
<td>0.397</td>
<td>9.498</td>
</tr>
<tr>
<td>Unnamed Sys by jspi (single model)</td>
<td>13.793</td>
<td>39.720</td>
<td>0.449</td>
<td>34.224</td>
<td>2.579</td>
<td>20.048</td>
</tr>
</tbody>
</table>

---


