A Chinese Span-Extraction Dataset for Machine Reading Comprehension

Abstract

This paper introduces a novel dataset for Chinese machine reading comprehension, focusing on span extraction. The data set is constructed using roughly 20,000 real-world questions that are annotated by experts on passages extracted from Wikipedia. A challenge set is also created with questions that demand a deep understanding and inference across multiple sentences. We also show several baseline models and anonymous submission scores to emphasize the challenges present in this dataset. The release of this dataset facilitated the Second Evaluation Workshop on Chinese Machine Reading Comprehension, also called CMRC 2018. We anticipate that this dataset will further facilitate research in Chinese machine reading comprehension.

1 Introduction

The capacity to interpret and comprehend natural language is a crucial component of achieving advanced artificial intelligence. Machine Reading Comprehension (MRC) is designed to understand the context of given texts and respond to related questions. Numerous types of MRC datasets have been developed, such as cloze-style reading comprehension, span-extraction reading comprehension, open-domain reading comprehension, and multiple-choice reading comprehension. Along with the increasing availability of reading comprehension datasets, several neural network methods have been proposed, leading to substantial advancements in this area.

There have also been various efforts to create Chinese machine reading comprehension datasets. In cloze-style reading comprehension, a Chinese cloze-style reading comprehension dataset was proposed, namely People's Daily Children's Fairy Tale. To increase the difficulty of the dataset, they also release a human-annotated evaluation set in addition to the automatically generated development and test sets. Later, another dataset was introduced using children's reading materials. To promote diversity and explore transfer learning, they also offer a human-annotated evaluation dataset using more natural queries compared to the cloze type. This dataset was the main component in the first evaluation workshop on Chinese machine reading comprehension (CMRC 2017). Furthermore, a large-scale open-domain Chinese machine reading comprehension dataset (DuReader) was created, containing 200k queries from search engine user query logs. There is also a reading comprehension dataset in Traditional Chinese.

While current machine learning techniques have outperformed human-level performance on datasets like SQuAD, it is still unclear whether similar results can be achieved on datasets using different languages. To accelerate the progress of machine reading comprehension research, we present a span-extraction dataset tailored for Chinese.

2 The Proposed Dataset

2.1 Task Definition

The reading comprehension task can be described as a triple (P, Q, A), where P is the passage, Q represents the question, and A is the answer. Specifically, in span-extraction reading comprehension, questions are created by humans which is a more natural way of creating data than the cloze-style MRC datasets. The answer A should consist of a specific span from the given passage P. The task can be simplified by predicting the start and end indices of the answer within the passage.

2.2 Data Pre-Processing

We downloaded the Chinese portion of Wikipedia from a specified date and used an open-source toolkit to process the raw files into plain text. Additionally, the Traditional Chinese characters were converted to Simplified Chinese to ensure consistency using another open-source tool.

2.3 Human Annotation

The questions in this dataset were created entirely by human experts, setting it apart from prior works that relied on automated data generation methods. Initially, documents are divided into passages, each containing no more than 500 Chinese words. Annotators are required to assess each passage for its suitability, discarding those that are too difficult for public understanding. Passages were discarded based on the following rules:

- If more than 30% of the passage consists of non-Chinese characters.
- If the passage includes too many specialized or professional terms.
- If the passage has a large number of special characters or symbols.
- If the paragraph is written in classical Chinese.

After determining that the passage is suitable, annotators generate questions and their corresponding primary answers based on the provided passage. During this question annotation, the following rules are used.

- Each passage should have no more than five questions.
- Answers must be a span from the passage.
- Question diversity is encouraged such as questions of type who, when, where, why, and how.
- Avoid copying descriptions from the passage directly. Use paraphrasing or syntax transformations to make answering more difficult.
- Long answers (over 30 characters) will be discarded.

For the evaluation sets, which include the development, test, and challenge sets, three answers are available for a more thorough assessment. Besides the primary answer generated by the question proposer, two additional annotators write a second and third answer for each question. These additional annotators do not see the primary answer to avoid biased answers.

2.4 Challenge Set

A challenge set was made to evaluate how effectively models can perform reasoning over diverse clues in the context, while still maintaining the span-extraction format. This annotation was also completed by three annotators. The questions in this set need to meet the following criteria:

- The answer can not be deduced from a single sentence in the passage if the answer is a single word or a short phrase. The annotation should encourage asking complex questions that need an overall view of the passage to answer correctly.
- If the answer is a named entity or belongs to a particular genre, it cannot be the only instance in the passage. There should be more than one instance to make the correct choice more difficult for the model.

2.5 Statistics

The overall statistics of the pre-processed data are shown in Table 1. The distribution of question types in the development set is shown in Figure 2.

Table 1: Statistics of the CMRC 2018 dataset.

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

3 Evaluation Metrics

This paper uses two evaluation metrics. Common punctuations and white spaces are ignored for normalization during evaluation.

3.1 Exact Match

The Exact Match (EM) score measures the exact overlap between the prediction and the ground truth answer. If the match is exact, then the score is 1; otherwise, the score is 0.

3.2 F1-Score

The F1-score evaluates the fuzzy overlap at the character level between the prediction and the ground truth answers. Instead of treating the answers as a bag of words, we calculate the longest common sequence (LCS) between the prediction and the ground truth and then compute the F1-score. The maximum F1 score among all the ground truth answers is taken for each question.

3.3 Estimated Human Performance

The estimated human performance is computed to measure the difficulty of the proposed dataset. Each question in the development, test, and challenge set has three answers. We use a cross-validation method to compute the performance. We treat the first answer as a human prediction and consider the other two answers as ground truth. Using this process, three human prediction scores are generated. Finally, we calculate the average of these three scores as the estimated human performance.

4 Experimental Results

4.1 Baseline System

We use BERT as the foundation of our baseline system. We modified the original script to accommodate our dataset. The initial learning rate was set to 3e-5, with a batch size of 32, and the training was conducted for two epochs. The document and query maximum lengths were set to 512 and 64 respectively.

4.2 Results

The results are in Table 2. Besides the baseline results, we include the results of the participants in the CMRC 2018 evaluation. The training and development sets were released to the public, and submissions were accepted to evaluate the models on the hidden test and challenge sets. As we can see that most of the participants achieved an F1 score above 80 in the test set. On the other hand, the EM metric shows considerably lower scores in comparison to the SQuAD dataset, highlighting that determining the precise span boundary is crucial for performance enhancement in Chinese reading comprehension.

As shown in the last column of Table 2, the top-ranked systems achieve decent results on the development and test sets but struggle to give satisfactory results on the challenge set. The estimated

Table 2: Baseline results and CMRC 2018 participants' results.

	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

human performance remains similar across the development, test, and challenge sets, indicating that the difficulty is consistent across all three data sets. Even though Z-Reader achieved the best performance on the test set, its EM metric performance was not consistent on the challenge set. This highlights that current models are limited in their ability to process difficult questions that require complex reasoning over numerous clues throughout the passage.

BERT-based methods demonstrated competitive performance compared to the submissions of participants. Traditional models have higher scores in the test set. However, the BERT-based models perform better on the challenge set, indicating the importance of rich representations to address complex questions.

5 Conclusion

This paper introduces a span-extraction dataset for Chinese machine reading comprehension, consisting of roughly 20,000 questions annotated by human experts, along with a challenge set which contains questions that need reasoning over different clues in the passage. The results from the evaluation suggest that models can achieve excellent scores on the development and test sets, close to the human performance in F1-score. However, the scores on the challenge set decline drastically, while human performance remains consistent. This shows there are still potential challenges in creating models that can perform well on difficult reasoning questions. We expect that this dataset will contribute to linguistic diversity in machine reading comprehension and facilitate additional research on questions that require comprehensive reasoning across multiple clues.