Investigating Prior Knowledge for Challenging Chinese Machine Reading Comprehension

Kai Sun1∗ Dian Yu2 Dong Yu2 Claire Cardie1

1Cornell University, Ithaca, NY
2Tencent AI Lab, Bellevue, WA
ks985@cornell.edu, {yudian, dyu}@tencent.com, cardie@cs.cornell.edu

Abstract

Machine reading comprehension tasks require a machine reader to answer questions relevant to the given document. In this paper, we present the first free-form multiple-choice Chinese machine reading Comprehension dataset (C3), containing 13,369 documents (dialogues or more formally written mixed-genre texts) and their associated 19,577 multiple-choice free-form questions collected from Chinese-as-a-second-language examinations.

We present a comprehensive analysis of the prior knowledge (i.e., linguistic, domain-specific, and general world knowledge) needed for these real-world problems. We implement rule-based and popular neural methods and find that there is still a significant performance gap between the best performing model (68.5%) and human readers (96.0%), especially on problems that require prior knowledge. We further study the effects of distractor plausibility and data augmentation based on translated relevant datasets for English on model performance. We expect C3 to present great challenges to existing systems as answering 86.8% of questions requires both knowledge within and beyond the accompanying document, and we hope that C3 can serve as a platform to study how to leverage various kinds of prior knowledge to better understand a given written or orally oriented text. C3 is available at https://dataset.org/c3/.

1 Introduction

“Language is, at best, a means of directing others to construct similar-thoughts from their own prior knowledge.”

Adams and Bruce (1982)

∗Part of this work was conducted when K. S. was an intern at the Tencent AI Lab, Bellevue, WA.
promoting the development of MRC techniques for Chinese, we introduce the first free-form multiple-Choice Chinese machine reading Comprehension dataset (C³) that not only contains multiple types of questions but also requires both the information in the given document and prior knowledge to answer questions. In particular, for assessing model generalizability across different domains, C³ includes a dialogue-based task C³D in which the given document is a dialogue, and a mixed-genre task C³M in which the given document is a mixed-genre text that is relatively formally written. All problems are collected from real-world Chinese-as-a-second-language examinations carefully designed by experts to test the reading comprehension abilities of language learners of Chinese.

We perform an in-depth analysis of what kinds of prior knowledge are needed for answering questions correctly in C³ and two representative free-form multiple-choice MRC datasets for English (Lai et al., 2017; Sun et al., 2019a), and to what extent. We find that solving these general-domain problems requires linguistic knowledge, domain-specific knowledge, and general world knowledge, the latter of which can be further broken down into eight types such as arithmetic, connotation, cause-effect, and implication. These free-form MRC datasets exhibit similar characteristics in that (i) they contain a high percentage (e.g., 86.8% in C³) of questions requiring knowledge gained from the accompanying document as well as at least one type of prior knowledge and (ii) regardless of language, dialogue-based MRC tasks tend to require more general world knowledge and less linguistic knowledge compared with tasks accompanied with relatively formally written texts. Specifically, compared with existing MRC datasets for Chinese (He et al., 2017; Cui et al. 2018b), C³ requires more general world knowledge (57.3% of questions) to arrive at the correct answer options.

We implement rule-based and popular neural approaches to the MRC task and find that there is still a significant performance gap between the best-performing model (68.5%) and human readers (96.0%), especially on problems that require prior knowledge. We find that the existence of wrong answer options that highly superficially match the given text plays a critical role in increasing the difficulty level of questions and the demand for prior knowledge. Furthermore, additionally introducing 94k training instances based on translated free-form multiple-choice datasets for English can only lead to a 4.6% improvement in accuracy, still far from closing the gap to human performance. Our hope is that C³ can serve as a platform for researchers interested in studying how to leverage different types of prior knowledge for in-depth text comprehension and facilitate future work on crosslingual and multilingual machine reading comprehension.

2 Related Work

Traditionally, MRC tasks have been designed to be text-dependent (Richardson et al., 2013; Hermann et al., 2015): They focus on evaluating comprehension of machine readers based on a given text, typically by requiring a model to answer questions relevant to the text. This is distinguished from many question answering (QA) tasks (Fader et al., 2014; Clark et al., 2016), in which no ground truth document supporting answers is provided with each question, making them relatively less suitable for isolating improvements to MRC. We will first discuss standard MRC datasets for English, followed by MRC/QA datasets for Chinese.

English. Much of the early MRC work focuses on designing questions whose answers are spans from the given documents (Hermann et al., 2015; Hill et al., 2016; Bajgar et al., 2016; Rajpurkar et al., 2016; Trischler et al., 2017; Joshi et al., 2017). As a question and its answer are usually in the same sentence, state-of-the-art methods (Devlin et al., 2019) have outperformed human performance on many such tasks. To increase task difficulty, researchers have explored a number of options including adding unanswerable (Trischler et al., 2017; Rajpurkar et al., 2018) or conversational (Choi et al., 2018; Reddy et al., 2019) questions that might require reasoning (Zhang et al., 2018a), and designing abstractive answers (Nguyen et al., 2016; Kočiský et al., 2018; Dalvi et al., 2018) or (question, answer) pairs that involve cross-sentence or cross-document content (Welbl et al., 2018; Yang et al., 2018). In general, most questions concern the facts that are explicitly expressed in the text, making these tasks possible to measure the
level of fundamental reading skills of machine readers.

Another research line has studied MRC tasks, usually in a free-form multiple-choice form, containing a significant percentage of questions that focus on the understanding of the implicitly expressed facts, events, opinions, or emotions in the given text (Richardson et al., 2013; Mostafazadeh et al., 2016; Khashabi et al., 2018; Lai et al., 2017; Sun et al., 2019a). Therefore, these benchmarks may allow a relatively comprehensive evaluation of different reading skills and require a machine reader to integrate prior knowledge with information presented in a text. In particular, real-world language exams are ideal sources for constructing this kind of MRC dataset as they are designed with a similar goal of measuring different reading comprehension abilities of human language learners primarily based on a given text. Representative datasets in this category include RACE (Lai et al., 2017) and DREAM (Sun et al., 2019a), both collected from English-as-a-foreign-language exams designed for Chinese learners of English. \( C^3_M \) and \( C^3_D \) can be regarded as a Chinese counterpart of RACE and DREAM, respectively, and we will discuss their similarities in detail in Section 3.3.

### Chinese

Extractive MRC datasets for Chinese (Cui et al., 2016; Li et al., 2016; Cui et al., 2018b; Cui et al., 2018a; Shao et al., 2018) have also been constructed—using web documents, news reports, books, and Wikipedia articles as source documents—and for which all answers are spans or sentences from the given documents. Zheng et al. (2019) propose a cloze-style multiple-choice MRC dataset by replacing idioms in a document with blank symbols, and the task is to predict the correct idiom from candidate idioms that are similar in meanings. The abstractive dataset DuReader (He et al., 2017) contains questions collected from query logs, free-form answers, and a small set of relevant texts retrieved from web pages per question. In contrast, \( C^3 \) is the first free-form multiple-choice Chinese MRC dataset that contains different types of questions and requires rich prior knowledge especially general world knowledge for a better understanding of the given text. Furthermore, 48.4% of problems require dialogue understanding, which has not been studied yet in existing Chinese MRC tasks.

Similarly, questions in many existing multiple-choice QA datasets for Chinese (Cheng et al., 2016; Guo et al., 2017a,b; Zhang and Zhao, 2018; Zhang et al., 2018b; Hao et al., 2019; Huang et al., 2019) are also free-form and collected from exams. However, most of the pre-existing QA tasks for Chinese are designed to test the acquisition and exploitation of domain-specific (e.g., history, medical, and geography) knowledge rather than general reading comprehension, and the performance of QA systems is partially dependent on the performance of information retrieval or the relevance of external resource (e.g., corpora or knowledge bases). We compare \( C^3 \) with relevant MRC/QA datasets for Chinese and English in Table 1.

### 3 Data

In this section, we describe the construction of \( C^3 \) (Section 3.1). We also analyze the data (Section 3.2) and the types of prior knowledge needed for the MRC tasks (Section 3.3).

#### 3.1 Collection Methodology and Task Definitions

We collect the general-domain problems from Hanyu Shuiping Kaoshi (HSK) and Minzu Hanyu Kaoshi (MHK), which are designed for evaluating the Chinese listening and reading comprehension ability of second-language learners such as international students, overseas Chinese, and ethnic minorities. We include problems from both real and practice exams; all are freely accessible online for public usage.

Each problem consists of a document and a series of questions. Each question is associated with several answer options, and exactly one of them is correct. The goal is to select the correct option. According to the document type, we divide these problems into two subtasks: \( C^3 \)-Dialogue (\( C^3_D \)), in which a dialogue serves as the document, and \( C^3 \)-Mixed (\( C^3_M \)), in which the given non-dialogue document is of mixed genre, such as a story, a news report, a monologue, or an advertisement. We show a sample problem for each type in Tables 2 and 3, respectively.

We remove duplicate problems and randomly split the data (13,369 documents and 19,577 questions in total) at the problem level, with 60% training, 20% development, and 20% test.
3.2 Data Statistics

We summarize the overall statistics of C³ in Table 4. We observe notable differences exist between C³M and C³. For example, C³M, in which most documents are formally written texts, has a larger vocabulary size compared to that of C³ with documents in spoken language. Similar observations have been made by Sun et al. (2019a) that the vocabulary size is relatively small in English dialogue-based machine reading comprehension tasks. In addition, the average document length (180.2) in C³M is longer than that in C³ (76.3). In general, C³ may not be suitable for evaluating the comprehension ability of machine readers on lengthy texts as the average length of document C³ is relatively short compared to that in datasets such as DuReader (He et al., 2017) (396.0) and RACE (Lai et al., 2017) (321.9).

3.3 Categories of Prior Knowledge

Previous studies on Chinese machine reading comprehension focus mainly on the linguistic knowledge required (He et al., 2017; Cui et al., 2018a). We aim instead for a more comprehensive analysis of the types of prior knowledge for answering questions. We carefully analyze a subset of questions randomly sampled from the development and test sets of C³ and arrive at the following three kinds of prior knowledge required for answering questions. A question is labeled as matching if it exactly matches or nearly matches (without considering determiners, aspect particles, or conjunctive adverbs; Xia, 2000) a span in the given document; answering questions in this category seldom requires any prior knowledge.

**LINGUISTIC:** To answer a given question (e.g., Q 1-2 in Table 2 and Q3 in Table 3), we require lexical/syntactic knowledge including but not limited to: idioms, proverbs, negation, antonymy, synonymy, the possible meanings of the word, and syntactic transformations (Nassaji, 2006).

**DOMAIN-SPECIFIC:** This kind of world knowledge consists of, but is not limited to, facts about domain-specific concepts, their definitions and properties, and relations among these concepts (Grishman et al., 1983; Hansen, 1994).

**GENERAL WORLD:** It refers to the general knowledge about how the world works, sometimes called commonsense knowledge. We focus on the sort of world knowledge that an encyclopedia would assume readers know without being told (Lenat et al., 1985; Schubert, 2002) instead of the factual knowledge such as properties of famous entities. We further break down general world knowledge into eight subtypes, some of which (marked with †) are similar to the categories summarized by LoBue and Yates (2011) for textual entailment recognition.

- **Arithmetic:** This includes numerical computation and analysis (e.g., comparison and unit conversion).
- **Connotation:** Answering questions requires knowledge about implicit and implied sentiment towards something or somebody, emotions, and tone (Edmonds and Hirst, 2002;...
In 1928, recommended by Hsu Chih-Mo, Hu Shih, who was the president of the previous National University of China, employed Shen Ts’un-g-wen as a lecturer of the university in charge of teaching the optional course of modern literature.

At that time, Shen was already famous in the literary world and was a little famous in society. Even before the beginning of class, the classroom was crowded with students. Upon the arrival of class, Shen went into the classroom. Seeing a dense crowd of students sitting beneath the platform, Shen was suddenly startled and his mind went blank. He was even unable to utter the first sentence he had rehearsed repeatedly.

He stood there motionlessly, extremely embarrassed. He wrung his hands without knowing where to put them. Before class, he had expected that he had a ready plan to meet the situation so he did not bring his teaching plan and textbook. For up to 10 minutes, the classroom was in perfect silence. All the students were curiously waiting for the new teacher to open his mouth. Breathing deeply, he gradually calmed down. Thereupon, the materials he had previously prepared gathered in his mind for the second time. Then he began his lecture. Nonetheless, since he was still nervous, it took him less than 15 minutes to finish the teaching contents he had planned to complete in an hour.

What should he do next? He was又一次 caught in embarrassment. He had no choice but to pick up a piece of chalk before writing several words on the blackboard: This is the first time I have given a lecture. In the presence of a crowd of people, I feel terrified.

Immediately, a peal of friendly laughter filled the classroom. Presently, a round of encouraging applause was given to him.

Q1 In paragraph 2, "a dense crowd" refers to
A. the light in the classroom was dim.
B. the number of students attending his lecture was large.
C. the room was noisy.
D. the students were active in voicing their opinions.

Q2 Shen did not bring the textbook because he felt that
A. the teaching contents were not many.
B. his preparation was sufficient.
C. his mental pressure could be reduced in this way.
D. the textbook was likely to restrict his ability to give a lecture.

Q3 Seeing the sentence written by Shen, the students
A. hurriedly consoled him.
B. blamed him in mind.
C. were greatly encouraged.
D. expressed their understanding and encouraged him.

Q4 The passage above is mainly about
A. the development of the Chinese educational system.
B. how to make self-adjustment if one is nervous.
C. the situation where Shen gave his lecture for the first time.
D. how Shen turned into a teacher from a writer.

---

Table 2: A C³ problem (left) and its English translation (right) (*: the correct option).

Feng et al., 2013; Van Hee et al., 2018). For example, the following conversation: 

"F: Ming Yu became a manager when he was so young! That’s impressive! M: It is indeed not easy!" is delivered in a tone for praise.

- **Cause-effect**: The occurrence of event A causes the occurrence of event B. We usually need this kind of knowledge to solve "why" questions when a causal explanation is not explicitly expressed in the given document.

- **Implication**: This category refers to the main points, suggestions, opinions, facts, or event predictions that are not expressed explicitly in the text, which cannot be reached by paraphrasing sentences using linguistic knowledge. For example, Q4 in Table 2 and Q2 in Table 3 belong to this category.

- **Part-whole**: We require knowledge that object A is a part of object B. Relations such as member-of, stuff-of, and component-of between two objects also fall into this category (Winston et al., 1987; Miller, 1998). For example, we require implication mentioned above as well as part-whole knowledge (i.e., "teacher" is a kind of job) to summarize the main topic of the following
Table 3: English translation of a sample problem from C³. Dialogue (C³_D) (*: the correct option).

| Question | Scenario | Dialogue (C³_D) |
|----------|----------|----------------|
| Q1 | What is the relationship between the speakers? | A. father and daughter. | B. mother and son. (**) | C. classmates. | D. colleagues. |
| Q2 | What difficulty has the male met? | A. his company does not have a vacation. | B. things are expensive during the Spring Festival. | C. he has not bought his ticket. | D. he cannot find the railway station. |
| Q3 | What suggestion does the male put forth? | A. he invites the female to come to Shanghai. (**) | B. he is going to line up in the next day. | C. he wants to go to the company as soon as possible. | D. he is going to go home after the Spring Festival is over. |

F: How is it going? Have you bought your ticket?
M: There are so many people at the railway station. I have waited in line all day long. However, when my turn comes, they say that there is no ticket left unless the Spring Festival is over.
F: It doesn’t matter. It is all the same for you to come back after the Spring Festival is over.
M: But according to our company’s regulation, I must go to the office on the 6th day of the first lunar month. I’m afraid I have no time to go back after the Spring Festival, so could you and my dad come to Shanghai for the coming Spring Festival?
F: I am too old to endure the travel.
M: It is not difficult at all. After I help you buy the tickets, you can come here directly.

Dialogue as ‘‘profession’’: ‘‘F: Many of my classmates become teachers after graduation. M: The best thing about being a teacher is feeling happy every day as you are surrounded by students!’’.

- Scenario: We require knowledge about observable behaviors or activities of humans and their corresponding temporal/locational information. We also need knowledge about personal information (e.g., profession, education level, personality, and mental or physical status) of the involved participant and relations between the involved participants, implicitly indicated by the behaviors or activities described in texts. For example, we put Q3 in Table 2 in this category as ‘‘friendly laughter’’ may express ‘‘understanding’’. Q1 in Table 3 about the relation between the two speakers also belongs to this category.

- Precondition: If event A had not happened, event B would not have happened (Ikuta et al., 2014; O’Gorman et al., 2016). Event A is usually mentioned in either the question or the correct answer option(s). For example, ‘‘I went to a supermarket’’ is a necessary precondition for ‘‘I was shopping at a supermarket when my friend visited me’’.

- Other: Knowledge that belongs to none of the above subcategories.

Two annotators (authors of this paper) annotate the type(s) of required knowledge for each question over 600 instances. To explore the differences and similarities in the required knowledge types between C³ and existing free-form MRC datasets, following the same annotation schema, we also annotate instances from the largest Chinese free-form abstractive MRC dataset DuReader (He et al., 2017) and free-form multiple-choice English MRC datasets RACE (Lai et al., 2017) and DREAM (Sun et al., 2019a), which can be regarded as the English counterpart of C³_M and C³_D, respectively. We also divide questions into one of three types—single, multiple, or independent—based on the minimum number of sentences in the document that explicitly or implicitly support the correct answer option. We regard a question as independent if it is context-independent, which usually requires prior vocabulary or domain-specific knowledge. The Cohen’s kappa coefficient is 0.62.

C³_M vs. C³_D As shown in Table 5, compared with the dialogue-based task (C³_D), C³_M with non-dialogue texts as documents requires more linguistic knowledge (49.0% vs. 30.7%) yet less general world knowledge (50.7% vs. 64.0%). As many as 24.3% questions in C³_D need scenario knowledge, perhaps due to the fact that speakers in a dialogue (especially face-to-face) may not explicitly mention information that they assume others already know such as personal information, the relationship between the speakers, and temporal and location information. Interestingly, we observe a similar phenomenon when we compare the English datasets DREAM (dialogue-based) and RACE. Therefore, it is likely that dialogue-based free-form tasks can serve as ideal platforms for studying how to improve language understanding with general world knowledge regardless of language.

C³ vs. its English counterparts We are also interested in whether answering a specific type of question may require similar types of prior knowledge across languages. For example, C³_D and its English counterpart DREAM (Sun et al., 2019a) have similar problem formats, document
Table 4: The overall statistics of $C^3$. $C^3 = C^3_M \cup C^3_D$.

| Metric | $C^3_M$ | $C^3_D$ | $C^3$ |
|--------|---------|---------|-------|
| Min./Avg./Max. # of options per question | 2 / 3.7 / 4 | 3 / 3.8 / 4 | 2 / 3.8 / 4 |
| # of correct options per question | 1 | 1 | 1 |
| Min./Avg./Max. # of questions per document | 1 / 1.9 / 6 | 1 / 1.2 / 6 | 1 / 1.5 / 6 |
| Avg./Max. question length (in characters) | 6.5 / 45 | 4.4 / 31 | 5.5 / 45 |
| Avg./Max. document length (in characters) | 13.5 / 57 | 10.9 / 34 | 12.2 / 57 |
| Avg./Max. document length (in characters) | 180.2 / 1,274 | 76.3 / 1,540 | 116.9 / 1,540 |
| Character vocabulary size | 4,120 | 2,922 | 4,193 |
| # of correct options per question | 1 | 1 | 1 |
| Min./Avg./Max. # of questions per document | 3 / 3.8 / 4 | 2 / 3.8 / 4 | 3 / 3.8 / 4 |
| Non-extractive correct option (%) | 81.9 | 78.9 | 80.4 |
| # of documents / # of questions | Training | 3,138 / 6,013 | 4,885 / 11,869 | 8,023 / 19,577 |
| Development | 1,046 / 1,991 | 1,628 / 1,825 | 2,674 / 3,816 |
| Test | 1,045 / 2,002 | 1,627 / 1,890 | 2,672 / 3,892 |
| All | 5,229 / 10,006 | 8,140 / 9,571 | 13,369 / 19,577 |

Table 5: Distribution (%) of types of required prior knowledge based on a subset of test and development sets of $C^3$, Chinese free-form abstractive dataset DuReader (He et al., 2017), and English free-form multiple-choice datasets RACE (Lai et al., 2017) and DREAM (Sun et al., 2019a). Answering a question may require more than one type of prior knowledge.

| Types | $C^3_M$ | $C^3_D$ | RACE | DREAM | DuReader |
|-------|---------|---------|------|-------|----------|
| Matching | 12.0 | 14.3 | 13.2 | 14.7 | 8.7 | 62.0 |
| Prior knowledge | 88.0 | 85.7 | 86.8 | 85.3 | 91.3 | 38.0 |
| - Linguistic | 49.0 | 30.7 | 39.8 | 47.3 | 40.0 | 22.0 |
| - Domain-specific | 0.7 | 1.0 | 0.8 | 0.0 | 0.0 | 16.0 |
| - General world | 50.7 | 64.0 | 57.3 | 43.3 | 57.3 | 0.0 |
| Arithmetic | 3.0 | 4.7 | 3.8 | 3.3 | 1.3 | 0.0 |
| Connotation | 1.3 | 5.3 | 3.3 | 3.3 | 5.3 | 0.0 |
| Cause-effect | 14.0 | 6.7 | 10.3 | 2.7 | 3.3 | 0.0 |
| Implication | 17.7 | 20.3 | 19.0 | 24.0 | 26.7 | 0.0 |
| Part-whole | 5.0 | 5.0 | 4.0 | 5.0 | 7.3 | 0.0 |
| Precondition | 2.7 | 4.3 | 3.5 | 2.7 | 1.3 | 0.0 |
| Scenario | 9.6 | 24.3 | 17.0 | 7.3 | 21.3 | 0.0 |
| Other | 3.3 | 0.3 | 1.8 | 2.0 | 0.7 | 0.0 |
| Single sentence | 50.7 | 22.7 | 36.7 | 24.0 | 12.0 | 14.6 |
| Multiple sentences | 47.0 | 77.0 | 62.0 | 75.3 | 88.0 | 68.7 |
| Independent | 2.3 | 0.3 | 1.3 | 0.7 | 0.0 | 16.7 |
| # of annotated instances | 300 | 300 | 600 | 150 | 150 | 150 |

C$^3$ vs. DuReader

The 150 annotated instances of DuReader also exhibit properties similar to those identified in studies of abstractive MRC for English (Nguyen et al., 2016; Kočiský et al., 2018; Reddy et al., 2019). Namely, turkers asked to write answers in their own words tend instead to write an extractive summary by copying short textual snippets or whole sentences in the given documents; this may explain why models designed for extractive MRC tasks achieve reasonable performance on abstractive tasks. We notice that questions in DuReader seldom require general world knowledge, which is possibly because users seldom ask questions about facts obvious to most people. On the other hand, as many as 16.7% of (question, answer) pairs in DuReader cannot be supported by the given text (vs. 1.3% in C$^3$); in most cases, they require prior knowledge about a particular domain (e.g., ‘‘On which website can I watch The Glory of Tang Dynasty?’’ and ‘‘How to start a clothing store?’’). In comparison, a larger fraction of C$^3$ requires linguistic knowledge or general world knowledge.
4 Approaches

We implement a classical rule-based method and recent state-of-the-art neural models.

4.1 Distance-Based Sliding Window

We implement Distance-based Sliding Window (Richardson et al., 2013), a rule-based method that chooses the answer option by taking into account (1) lexical similarity between a statement (i.e., a question and an answer option) and the given document with a fixed window size and (2) the minimum number of tokens between occurrences of the question and occurrences of an answer option in the document. This method assumes that a statement is more likely to be correct if there is a shorter distance between tokens within a statement, and more informative tokens in the statement appear in the document.

4.2 Co-Matching

We utilize Co-Matching (Wang et al., 2018), a Bi-LSTM-based model for multiple-choice MRC tasks for English. It explicitly treats a question and one of its associated answer options as two sequences and jointly models whether or not the given document matches them. We modify the pre-processing step and adapt this model to MRC tasks for Chinese (Section 4.3).

4.3 Fine-Tuning Pre-Trained Language Models

We also apply the framework of fine-tuning a pre-trained language model on machine reading comprehension tasks (Radford et al., 2018). We consider the following four pre-trained language models for Chinese: Chinese BERT-Base (denoted as BERT) (Devlin et al., 2019), Chinese ERNIE-Base (denoted as ERNIE) (Sun et al., 2019b), and Chinese BERT-Base with whole word masking during pre-training (denoted as BERT-wwm) (Cui et al., 2019) and its enhanced version pre-trained over larger corpora (denoted as BERT-wwm-ext). These models have the same number of layers, hidden units, and attention heads.

Given document $d$, question $q$, and answer option $o_i$, we construct the input sequence by concatenating $[CLS]$, tokens in $d$, $[SEP]$, tokens in $q$, $[SEP]$, tokens in $o_i$, and $[SEP]$, where $[CLS]$ and $[SEP]$ are the classifier token and sentence separator in a pre-trained language model, respectively. We add an embedding vector $t_1$ to each token before the first $[SEP]$ (inclusive) and an embedding vector $t_2$ to every other token, where $t_1$ and $t_2$ are learned during language model pre-training for discriminating sequences. We denote the final hidden state for the first token in the input sequence as $S_i \in \mathbb{R}^{1 \times H}$, where $H$ is the hidden size. We introduce a classification layer $W \in \mathbb{R}^{1 \times H}$ and obtain the unnormalized log probability $P_i \in \mathbb{R}$ of $o_i$ being correct by $P_i = S_i W^T$. We obtain the final prediction for $q$ by applying a softmax layer over the unnormalized log probabilities of all options associated with $q$.

5 Experiment

5.1 Experimental Settings

We use $C^3_M$ and $C^3_D$ together to train a neural model and perform testing on them separately, following the default setting on RACE that also contains two subsets (Lai et al., 2017). We run every experiment five times with different random seeds and report the best development set performance and its corresponding test set performance.

Distance-Based Sliding Window. We simply treat each character as a token. We do not use Chinese word segmentation as it results in drops in performance based on our experiment.

Co-Matching. We replace the English tokenizer with a Chinese word segmenter in HanLP.¹ We use the 300-dimensional Chinese word embeddings released by Li et al. (2018).

Fine-Tuning Pre-Trained Language Models.

We set the learning rate, batch size, and maximal sequence length to $2 \times 10^{-5}$, 24, and 512, respectively. We truncate the longest sequence among $d$, $q$, and $o_i$ (Section 4.3) when an input sequence exceeds the length limit 512. For all experiments, we fine-tune a model on $C^3$ for eight epochs. We keep the default values for the other hyperparameters (Devlin et al., 2019).

5.2 Baseline Results

As shown in Table 6, methods based on pre-trained language models (BERT, ERNIE, BERT-wwm, and BERT-wwm-ext) outperform the Distance-based Sliding Window approach and Bi-LSTM-based Co-Matching by a large margin. BERT-wwm-ext performs better on $C^3$ compared

¹https://github.com/hankcs/HanLP.
with other three pre-trained language models, though there still exists a large gap (27.5%) between this method and human performance (96.0%).

We also report the performance of Co-Matching, BERT, BERT-wwm-ext, and human on different question categories based on the annotated development sets (Table 7), which consist of 150 questions in $C^3_M$ and 150 questions in $C^3_D$. These models generally perform worse on questions that require prior knowledge or reasoning over multiple sentences than questions that can be answered by surface matching or only need the information from a single sentence (Section 3.3).

### 5.3 Discussions on Distractor Plausibility

We look into incorrect predictions of Co-Matching, BERT, and BERT-wwm-ext on the development set. We observe that the existence of plausible distractors may play a critical role in raising the difficulty level of questions for models. We regard a distractor (i.e., wrong answer option) as plausible if it, compared with the correct answer option, is more superficially similar to the given document. Two typical cases include (1) the information in the distractor is accurate based on the document but does not (fully) answer the question, and (2) the distractor distorts, oversimplifies, exaggerates, or misinterprets the information in the document.

Given document $d$, the correct answer option $c$, and wrong answer options $\{w_1, w_2, \ldots, w_i, \ldots, w_n\}$ associated with a certain question, we measure the distractor plausibility of distractor $w_i$ by:

$$\gamma_i = S(w_i, d) - S(c, d)$$  \hspace{1cm} (1)
where $S(x, y)$ is a normalized similarity score between 0 and 1 that measures the edit distance to change $x$ into a substring of $y$ using single-character edits (insertions, deletions or substitutions). Particularly, if $x$ is a substring of $y$, $S(x, y) = 1$; if $x$ shares no character with $y$, $S(x, y) = 0$. By definition, $S(w_i, d)$ in Equation (1) measures the lexical similarity between distractor $w_i$ and $d$; $S(c, d)$ measures the similarity between the correct answer option $c$ and $d$.

To quantitatively investigate the impact of the existence of plausible distractors on model performance, we group questions from the development set of C$^3$ by the largest distractor plausibility (i.e., $\max_i \gamma_i$), in the range of [-1, 1], for each question and compare the performance of Co-Matching, BERT, and BERT-wwm-ext in different groups.

As shown in Figure 1(a), the largest distractor plausibility may serve as an indicator of the difficulty level of questions presented to the investigated models. When the largest distractor plausibility is smaller than -0.8, all three models exhibit strong performance ($\geq$ 90%). As the largest distractor plausibility increases, the performance of all models consistently drops. All models perform worse than average on questions having at least one high-plausible distractor (e.g., distractor plausibility > 0). Compared with BERT, the gain of the best-performing model (i.e., BERT-wwm-ext) mainly comes from its superior performance on these “difficult” questions.

Further, we find that distractor plausibility is strongly correlated with the need for prior knowledge when answering questions in C$^3$ based...
Figure 3: Performance of BERT-wwm-ext trained on 1/8, 2/8, ..., 8/8 of C3 training data, and C3 training data plus 1/8, 2/8, ..., 8/8 of machine translated (MT) RACE and DREAM training data.

Figure 3: Performance of BERT-wwm-ext trained on 1/8, 2/8, ..., 8/8 of C3 training data, and C3 training data plus 1/8, 2/8, ..., 8/8 of machine translated (MT) RACE and DREAM training data.

5.4 Discussions on Data Augmentation

To extrapolate to what extent we can improve the performance of current models with more training data, we plot the development set performance of BERT-wwm-ext trained on different portions of the training data of C3. As shown in Figure 3, the accuracy grows roughly linearly with the logarithm of the size of training data, and we observe a substantial gap between human performance and the expected BERT-wwm-ext performance, even assuming that 10^5 training instances are available, leaving much room for improvement.

Furthermore, as the knowledge type distributions of C3 and its English counterparts RACE and DREAM are highly similar (Section 3.3), we translate RACE and DREAM from English to Chinese with Google Translate and plot the performance of BERT-wwm-ext trained on C3 plus different numbers of translated instances. The learning curve is also roughly linear with the logarithm of the number of training instances from translated RACE and DREAM, but with a lower growth rate. Even augmenting the training data with all 94k translated instances only leads to a 4.6% improvement (from 67.8% to 72.4%) in accuracy on the development set of C3. From another perspective, BERT-wwm-ext trained on all translated instances without using any data in C3 only achieves an accuracy of 67.1% on the development set of C3, slightly worse than 67.8% achieved when only the training data in C3 is used, whose size is roughly 1/8 of that of the translated instances. These observations suggest a need to better leverage large-scale English resources from similar MRC tasks.

Besides augmenting the training data with translated instances, we also attempt to fine-tune a pre-trained multilingual BERT-Base released by Devlin et al. (2019) on the training data of C3 and all original training instances in English from RACE and DREAM. However, the accuracy on the development set of C3 is 63.4%, which is even lower than the performance (65.7% in Table 6) of fine-tuning Chinese BERT-Base only on C3.

6 Conclusion

We present the first free-form multiple-choice Chinese machine reading comprehension dataset (C3), collected from real-world language exams, requiring linguistic, domain-specific, or general world knowledge to answer questions based on the given written or orally oriented texts. We study the prior knowledge needed in this challenging machine reading comprehension dataset and carefully investigate the impacts of distractor plausibility and data augmentation (based on similar resources for English) on the performance of state-of-the-art neural models. Experimental results demonstrate the there is still a significant performance gap between the best-performing model (68.5%) and human readers (96.0%) and a need for better ways for exploiting rich resources in other languages.

Acknowledgments

We would like to thank the editors and anonymous reviewers for their helpful and insightful feedback.
References

Marilyn Adams and Bertram Bruce. 1982. Background knowledge and reading comprehension. Reader Meets Author: Bridging the Gap, 13:2–25.

Ondrej Bajgar, Rudolf Kadlec, and Jan Kleindienst. 2016. Embracing data abundance: Booktest dataset for reading comprehension. arXiv preprint, cs.CL/1610.00956v1.

Gong Cheng, Weixi Zhu, Ziwei Wang, Jianghui Chen, and Yuzhong Qu. 2016. Taking up the gaokao challenge: An information retrieval approach. In Proceedings of the IJCAI, pages 2479–2485. New York, NY.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In Proceedings of the EMNLP, pages 2174–2184. Brussels, Belgium.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Proceedings of the NAACL-HLT, pages 2924–2936. Minneapolis, MN.

Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter D. Turney, and Daniel Khashabi. 2016. Combining retrieval, statistics, and inference to answer elementary science questions. In Proceedings of the AAAI, pages 2580–2586. Phoenix, AZ.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for Chinese BERT. arXiv preprint, cs.CL/1906.08101v1.

Yiming Cui, Ting Liu, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2018a. Dataset for the first evaluation on chinese machine reading comprehension. In Proceedings of the LREC, pages 2721–2725. Miyazaki, Japan.

Yiming Cui, Ting Liu, Zhipeng Chen, Shijin Wang, and Guoping Hu. 2016. Consensus attention-based neural networks for Chinese reading comprehension. In Proceedings of the COLING, pages 1777–1786. Osaka, Japan.

Yiming Cui, Ting Liu, Li Xiao, Zhipeng Chen, Wentao Ma, Wanxiang Che, Shijin Wang, and Guoping Hu. 2018b. A span-extraction dataset for Chinese machine reading comprehension. In Proceedings of the EMNLP, pages 5882–5888. Hong Kong, China.

Bhavana Dalvi, Lifu Huang, Niket Tandon, Wen-tau Yih, and Peter Clark. 2018. Tracking state changes in procedural text: A challenge dataset and models for process paragraph comprehension. In Proceedings of the NAACL-HLT, pages 1595–1604. New Orleans, LA.

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 NAACL-HLT, pages 4171–4186. Minneapolis, MN.

Philip Edmonds and Graeme Hirst. 2002. Near-synonymy and lexical choice. Computational Linguistics, 28(2):105–144.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2014. Open question answering over curated and extracted knowledge bases. In Proceedings of the SIGKDD, pages 1156–1165. New York City, NY.

Song Feng, Jun Seok Kang, Polina Kuznetsova, and Yejin Choi. 2013. Connotation lexicon: A dash of sentiment beneath the surface meaning. In Proceedings of the ACL, pages1774–1784. Sofia, Bulgaria.

Ralph Grishman, Lynette Hirschman, and Carol Friedman. 1983. Isolating domain dependencies in natural language interfaces. In Proceedings of the ANLP, pages 46–53. Santa Monica, CA.

Shangmin Guo, Kang Liu, Shizhu He, Cao Liu, Jun Zhao, and Zhuoyu Wei. 2017a. IJCNLP-2017 Task 5: Multi-choice question answering in examinations. In Proceedings of the IJCNLP 2017, Shared Tasks, pages 34–40. Taipei, Taiwan.

Shangmin Guo, Xiangrong Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2017b. Which is the effective way for Gaokao: Information retrieval or neural networks? In Proceedings of the EACL, pages 111–120. Valencia, Spain.

Steffen Leo Hansen. 1994. Reasoning with a domain model. In Proceedings of the NODALIDA, pages 111–121. Stockholm, Sweden.
Yu Hao, Xien Liu, Ji Wu, and Ping Lv. 2019. Exploiting sentence embedding for medical question answering. In Proceedings of the AAAI, pages 938–945. Honolulu, HI.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2017. DuReader: a Chinese machine reading comprehension dataset from real-world applications. In Proceedings of the MRQA, pages 37–46. Melbourne, Australia.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In Proceedings of the NIPS, pages 1693–1701. Montreal, Canada.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children’s books with explicit memory representations. In Proceedings of the ICLR. San Juan, Puerto Rico.

Zixian Huang, Yulin Shen, Xiao Li, Yuang Wei, Gong Cheng, Lin Zhou, Xinyu Dai, and Yuzhong Qu. 2019. GeoSQA: A benchmark for scenario-based question answering in the geography domain at high school level. In Proceedings of the EMNLP-IJCNLP, pages 5865–5870. Hong Kong, China.

Rei Ikuta, Will Styler, Mariah Hamang, Tim O’Gorman, and Martha Palmer. 2014. Challenges of adding causation to richer event descriptions. In Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation, pages 12–20. Baltimore, MD.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. arXiv preprint, cs.CL/1705.03551v2.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In Proceedings of the NAACL-HLT, pages 252–262. New Orleans, LA.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gáabor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. Transactions of the Association of Computational Linguistics, 6:317–328.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale reading comprehension dataset from examinations. In Proceedings of the EMNLP, pages 785–794. Copenhagen, Denmark.

Adam Lally, Sugato Bagchi, Michael A. Barborak, David W. Buchanan, Jennifer Chu-Carroll, David A. Ferrucci, Michael R. Glass, Aditya Kalyanpur, Erik T. Mueller, J. William Murdock, Siddharth Patwardhan, and John M. Prager. 2017. WatsonPaths: Scenario-based question answering and inference over unstructured information. AI Magazine, 38(2):59–76.

Douglas B. Lenat, Mayank Prakash, and Mary Shepherd. 1985. CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. AI Magazine, 6(4):65–65.

Peng Li, Wei Li, Zhengyan He, Xuguang Wang, Ying Cao, Jie Zhou, and Wei Xu. 2016. Dataset and neural recurrent sequence labeling model for open-domain factoid question answering. arXiv preprint, cs.CL/1607.06275v2.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on chinese morphological and semantic relations. In Proceedings of the ACL, pages 138–143. Melbourne, Australia.

Peter LoBue and Alexander Yates. 2011. Types of common-sense knowledge needed for recognizing textual entailment. In Proceedings of the ACL, pages 329–334. Portland, OR.

George Miller. 1998. WordNet: An Electronic Lexical database. MIT Press.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. In Proceedings of the NAACL-HLT, pages 839–849. San Diego, CA.
Hossein Nassaji. 2006. The relationship between depth of vocabulary knowledge and L2 learners lexical inferencing strategy use and success. *The Modern Language Journal*, 90(3):387–401.

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*, cs.CL/1611.09268v3.

Tim O’Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the CNS*, pages 47–56. Austin, TX.

Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2018. SemEval-2018 Task 11: Machine comprehension using commonsense knowledge. In *Proceedings of the SemEval*, pages 747–757. New Orleans, LA.

Hoifung Poon, Janara Christensen, Pedro Domingos, Oren Etzioni, Raphael Hoffmann, Chloe Kiddon, Thomas Lin, Xiao Ling, Mausam, Alan Ritter, Stefan Schoenmackers, Stephen Soderland, Dan Weld, Fei Wu, and Congle Zhang. 2010. Machine reading at the University of Washington. In *Proceedings of the NAACL-HLT FAM-LbR*, pages 87–95. Los Angeles, CA.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. https://openai.com/blog/language-unsupervised/.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don’t know: Unanswerable questions for squad. In *Proceedings of the ACL*, pages 784–789. Melbourne, Australia.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the EMNLP*, pages 2383–2392. Austin, TX.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Matthew Richardson, Christopher J. C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the EMNLP*, pages 193–203. Seattle, WA.

Lenhart Schubert. 2002. Can we derive general world knowledge from texts? In *Proceedings of the HLT*, pages 94–97. San Diego, CA.

Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2018. DRCD: A Chinese machine reading comprehension dataset. *arXiv preprint*, cs.CL/1806.00920v3.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019a. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Yu Sun, Shuhuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019b. ERNIE: Enhanced representation through knowledge integration. *arXiv preprint*, cs.CL/1904.09223v1.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the RepL4NLP*, pages 191–200. Vancouver, Canada.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. We usually dont like going to the dentist: Using common sense to detect irony on twitter. *Computational Linguistics*, 44(4):793–832.

Shuohang Wang, Mo Yu, Shiyu Chang, and Jing Jiang. 2018. A co-matching model for multi-choice reading comprehension. In *Proceedings of the ACL*, pages 1–6. Melbourne, Australia.

Johannes Welbl, Pontus Stanetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association of Computational Linguistics*, 6:287–302.

Morton E. Winston, Roger Chaffin, and Douglas Herrmann. 1987. A taxonomy of part-whole relations. *Cognitive Science*, 11(4):417–444.
Fei Xia. 2000. The part-of-speech tagging guidelines for the Penn Chinese Treebank (3.0). IRCS Technical Reports Series, 1–43.

Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. 2018. Large-scale cloze test dataset created by teachers. In Proceedings of the EMNLP, pages 234–2356. Brussels, Belgium.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Proceedings of the EMNLP, pages 2369–2380. Brussels, Belgium.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018a. ReCoRD: Bridging the gap between human and machine commonsense reading comprehension. arXiv preprint, cs.CL/1810.12885v1.

Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. 2018b. Medical exam question answering with large-scale reading comprehension. In Proceedings of the AAAI, pages 5706–5713. New Orleans, LA.

Zhuosheng Zhang and Hai Zhao. 2018. One-shot learning for question-answering in Gaokao history challenge. In Proceedings of the COLING, pages 449–461. Santa Fe, NM.

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. ChID: A large-scale Chinese idiom dataset for cloze test. In Proceedings of the ACL, pages 778–787. Florence, Italy.