MACHINE READING COMPREHENSION: CHALLENGES AND APPROACHES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Kai Sun

August 2021

MACHINE READING COMPREHENSION: CHALLENGES AND APPROACHES

Kai Sun, Ph.D.

Cornell University 2021

Machine reading comprehension (MRC) tasks have attracted substantial attention from both academia and industry. These tasks require a machine reader to answer questions relevant to a given document provided as input. In this dissertation, we mainly focus on *non-extractive MRC*, in which a significant percentage of candidate answers are not restricted to text spans from the reference document or corpus. In comparison to *extractive MRC* tasks, non-extractive MRC tasks contain a significant percentage of questions focusing on the **implicitly** expressed facts, events, opinions, or emotions in the given text, requiring diverse types of world knowledge (e.g., commonsense, paraphrase, and arithmetic knowledge) and advanced reading skills (e.g., logical reasoning, summarization, and sentiment analysis). This dissertation presents our work in exploring new challenges and approaches for non-extractive MRC. Specifically, on the challenge side, we create the first MRC dataset that focuses on in-depth multi-turn multi-party dialogue understanding and the first free-form multiple-choice Chinese MRC dataset that requires various kinds of prior knowledge. On the approach side, we propose three general reading strategies and a method of utilizing contextualized knowledge to improve non-extractive MRC. We find our datasets to be very challenging for reading comprehension systems and our approaches to be empirically effective on representative non-extractive MRC tasks.

BIOGRAPHICAL SKETCH

Kai Sun was born and grew up in Shandong, China. He developed interests in Computer Science at an early age and later entered Shanghai Jiao Tong University to study Computer Science. During his undergraduate study, he started conducting research on speech and language processing with Professor Kai Yu. After obtaining his bachelor's degree, he came to Cornell University to pursue a doctorate degree in Computer Science, where he studied Natural Language Processing under the guidance of Professor Claire Cardie. He also did research internships at Microsoft Research Asia, Tencent AI Lab, and Facebook AI, working on various speech and language processing projects. Apart from work, he is actively involved in the computer gomoku and renju community. He designed *Yixin*, the first gomoku and renju AI program that can compete at the human champion level.

To my parents and Hui, for their unconditional love.

ACKNOWLEDGEMENTS

First and foremost, I am extremely grateful to my advisor, Claire Cardie, for her guidance, insights, suggestions, encouragement, time, patience, and help throughout my Ph.D. life. Claire is an incredibly kind and supportive advisor who always believes in me and gives me complete freedom in pursuing my interests. I especially want to thank her for supporting me to spend quite a lot of time working on computer gomoku and renju, an area I was passionate about but irrelevant to my primary Ph.D. research. I am also thankful to my minor advisors John Hopcroft and Thorsten Joachims, for their advice and feedback.

Furthermore, I am very grateful to my internship mentors: Dian Yu, Jianshu Chen, and Dong Yu (Tencent), Shane Moon, Paul Crook, Stephen Roller, Bing Liu, Stephen Wang, and Honglei Liu (Facebook). I especially want to say special thanks to Dian, who is my fantastic teammate, but more importantly, my very good friend. You were there in the trenches working with me on all the challenges tackled by this dissertation. I am so grateful to have had the opportunity to work with someone as wonderful as you. There are also many other collaborators and colleagues that contributed to my internship experience: Xiaoman Pan, Hai Wang, Yejin Choi, Guoyin Wang, Chengzhu Yu, Chao Weng, Becka Silvert, and Eunjoon Cho.

I would also like to thank Xilun Chen, Vlad Niculae, Yao Cheng, Rishi Bommasani, Arzoo Katiyar, Esin Durmus, Xinya Du, Menglin Jia, Tianze Shi, Felix Wu, Xanda Schofield, Ana Smith, Maria Antoniak, Ashudeep Singh, Jonathan Chang, Liye Fu, Chenhao Tan, Forrest Davis, Noriyuki Kojima, Lillian Lee, Yoav Artzi, Sasha Rush, Cristian Danescu-Niculescu-Mizil, and all other members of the Cornell NLP seminar for fun discussions and helpful feedback.

I want to thank Kai Yu, my Bachelor's thesis advisor, for a lot of guidance

and support in my early research. I also thank my other collaborators at SJTU SpeechLab, Qizhe Xie, Lu Chen, Su Zhu, Siqiu Yao, and Xueyang Wu, with whom I finished my dialogue state tracking research during my first Ph.D. year.

Computer gomoku and renju occupied a significant role in my Ph.D. journey. I have been lucky to be surrounded by great friends and supported by the whole community. I want to thank Tianyi Hao, Ming Lu, Qichao Wang, Rong Xiao, Runzhe Yang, Kai Yu, Tao Tao, and all other people who helped the development of Yixin. I thank Tianyi Hao, Tao Tao, Alexander Bogatirev, Epifanov Dmitry, Shu-Hsuan Lin, Rudolf Dupszki, Guan Qi, Xiaohan Dai, Makarov Pavel, Nikonov Konstantin, Kai Yu, and all other people and sponsors who contributed to the organization of formal matches between Yixin and top human players. I also want to thank Tianyi Hao, Tomas Kubes, Petr Lastovicka, and all other contributors of Gomocup.

Finally, I am deeply grateful to my parents for their love and long-lasting support. I am also immensely thankful to Hui for all the love and happiness she brought into my life, which completely changed my life and made it infinitely more colorful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Motivation

Natural Language Understanding (NLU) is one of the most challenging areas in the field of Artificial Intelligence (AI) (Shapiro, 1992). This dissertation tackles a central problem in NLU: how to teach machines to read a document in a human language and answer comprehension questions, i.e., machine reading comprehension (MRC). The problem of MRC is important, not only because the MRC tasks can help the research community measure the progress of NLU and general AI but also because the MRC techniques are essential in real-world applications such as building AI-powered virtual assistants.

There are many MRC tasks and various ways of categorization. At a high level, we can divide MRC tasks into three categories based on the answer type: *extractive*, *abstractive*, and *multiple-choice*.

(a) **Extractive MRC** refers to the task where the answer is a short span that can be extracted from the document. Table 1.1 (a) demonstrates an example. A major limitation of extractive MRC is that, for many questions, answers cannot be represented as a span of the document.

(b) In response to the limitation of extractive MRC, answers in **abstractive MRC** tasks are human-generated texts, which do not have to be spans of the document, as shown in Table 1.1 (b). However, the evaluation of abstractive MRC is non-trivial due to the variance of possible answers. Moreover, since annotators tend to copy spans as answers directly, the

(a) Extractive MRC
document: In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail ... **question:** What is another main form of precipitation besides drizzle, rain, snow, sleet and hail? **answer:** graupel
(b) Abstractive MRC
document: New Jersey is a state in the Northeastern and mid-Atlantic regions of the United States. It is a peninsula, bordered on the north and east by the state of New York ... **question:** Where is New Jersey located? **answer:** In the Northeastern and mid-Atlantic regions of the US
(c) Multiple-Choice MRC
document: Dirk Diggler was born as Steven Samuel Adams on April 15, 1961 outside of Saint Paul, Minnesota ... He was discovered at a falafel stand by Jack Horner. Diggler met his friend, Reed Rothchild, through Horner in 1979 while working on a film ... **question:** How old was Dirk when he met his friend Reed? **options:** A. 18 B. 16 C. 17 D. 15 **answer:** A

Table 1.1: Examples adapted from representation MRC tasks: (a) SQuAD (Rajpurkar et al., 2016), (b) CoQA (Reddy et al., 2019), and (c) MultiRC (Khashabi et al., 2018). Clues to the answers are underlined.

majority of answers are still extractive in many of these tasks (Kočiskỳ et al., 2018; Reddy et al., 2019).

(c) In **multiple-choice MRC**, multiple answer options are provided along with a question, and the goal is to choose the correct option(s). Table 1.1 (c) shows an example. Compared with abstractive MRC, we can adopt objective evaluation criteria such as accuracy to evaluate system performance more easily (Clark et al., 2016; Lai et al., 2017).

In this dissertation, we primarily focus on ***non-extractive*** multiple-choice MRC tasks, in which a significant percentage of answer options are not extractive text spans. Compared to questions in extractive MRC tasks, besides surface matching, there are various types of complicated questions such as math word problems, summarization, logical reasoning, and sentiment analysis, requiring

document: How quickly can you count from one to ten? Do you use ten different words to do it? Can you do it in English, or do you have to use your first language? Do you count on your fingers? Many people think that numbers and math are the same all over the world. But scientists have discovered that it is not true. People in different parts of the world use different ways to count on their fingers. In the United States, people begin counting with their first finger, which they extend or stick out. They then extend the rest of their fingers and finally the thumb to count to five. Then they repeat this with the other hand to get to ten. In China, people count by using different finger positions. In this way, a Chinese person can easily count to ten on only one hand. Besides ways of finger counting, scientists have found that cultures and languages are also different when it comes to numbers. Some languages have only a few words for numbers, and others have no words for numbers. A group of scientists studied aboriginal people in Australia. There people don't have hand movements to stand for numbers. They don't even have words for numbers. However, they are still able to understand different ideas about numbers. In a similar study, researchers from the Massachusetts Institute of Technology discovered that people of the Piraha tribe in northwestern Brazil don't have words for numbers such as "one" or "three". They are not able to say "five trees" or "ten trees" but can say "some trees", "more trees", or "many trees". Professor Edward Gibson said that most people believe that everyone knows how to count," but here is a group that does not count. They could learn, but it's not useful in their culture, so they've never picked it up." Although all humans are able to understand quantities, not all languages have numbers and not all people use counting. Number words in a certain language are a result of people needing numbers in their daily lives. Now we know that people have different ideas about numbers and math, too.
question 1: The writer begins with the four questions in order to _ .
options: A. make a survey B. interest readers C. tell a story D. solve math problems
answer: B
question 2: What is the main idea of the passage?
options: A. People from different cultures have different ideas about numbers and math.
B. Chinese people can count more easily on their fingers than Americans.
C. In some aboriginal cultures, people don't even know how to count.
D. Some languages don't have number words because people don't need numbers.
answer: A

Table 1.2: Examples of various types of questions from a non-extractive MRC dataset RACE (Lai et al., 2017).

advanced reading skills and prior world knowledge. For example, answering the question in Table 1.1 (c) requires math knowledge; answering questions in Table 1.2 requires commonsense knowledge and summarization ability. Due to the complexity, the problem of answering such challenging non-extractive questions has not been studied systematically until the 2010s (Richardson et al., 2013; Lai et al., 2017), despite that the study of MRC dates back to the 1970s (Charniak, 1972; Lehnert, 1977; Chen, 2018). Since 2017, the problem has received much attention, and rapid progress has been made, including our efforts that we present in this dissertation. See the progress on RACE (Lai et al., 2017), a representative non-extractive multiple-choice MRC dataset in Figure 1.1 as an example.

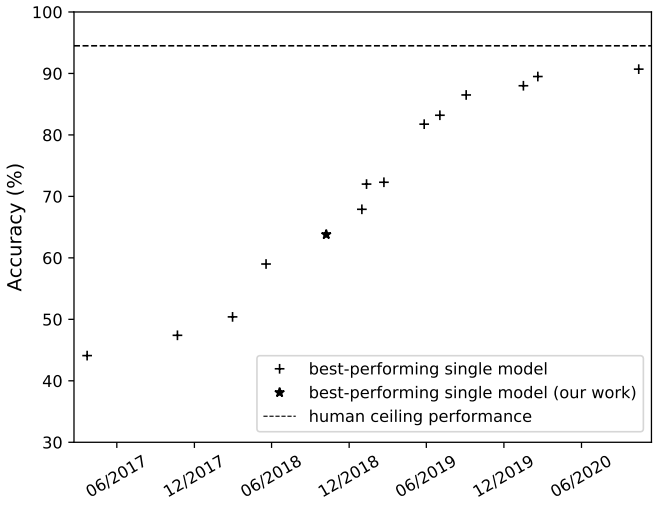


Figure 1.1: The best results from the official leaderboard of the RACE dataset (Lai et al., 2017) since its release in April 2017.

The progress in this field is mainly attributed to (i) the creation of a growing number of non-extractive MRC datasets and (ii) the development of various machine learning-based approaches for non-extractive MRC. This dissertation presents our efforts in both aspects. In terms of (i), we construct high-quality human-labeled datasets to support the study of two underexplored directions: dialogue comprehension and Chinese comprehension. As for (ii), we develop three general reading strategies and a method of utilizing contextualized knowledge to improve non-extractive MRC. We describe the corresponding contributions in more detail in the next section.

1.2 Contributions

Reading strategies for machine reading comprehension. Inspired by reading strategies identified in cognitive science, we propose three domain-independent reading strategies aimed to improve non-extractive machine reading comprehen-

sion. Our proposed strategies lead to substantial performance improvement over previous best results on seven representative non-extractive MRC tasks from different domains. We describe this work in detail in Chapter 3.

Dialogue-based machine reading comprehension. We construct the first dialogue-based multiple-choice reading comprehension dataset. In contrast to previous reading comprehension datasets whose source documents are generally drawn from formal written texts, our work is the first to focus on in-depth multi-turn multi-party dialogue understanding, which is likely to present significant challenges for reading comprehension systems. We investigate the effects of incorporating general world knowledge and dialogue structure into rule-based and machine learning-based MRC models and show the effectiveness of these factors, suggesting a promising direction for dialogue-based reading comprehension. We present this work in Chapter 4.

Chinese machine reading comprehension. To promote the development of MRC techniques for Chinese, we introduce the first free-form multiple-choice Chinese machine reading comprehension dataset that requires knowledge gained from the given document as well as prior knowledge to answer questions. We present a comprehensive analysis of the prior knowledge (i.e., linguistic, domain-specific, and general world knowledge) needed in this challenging dataset and study the effects of distractor plausibility and data augmentation based on translated relevant datasets for English on model performance. This work is the subject of Chapter 5.

Contextualized knowledge for machine reading comprehension. We explore the influence of verbal-nonverbal knowledge on MRC tasks, especially those that require tacit general world knowledge. We focus on interrelated verbal-nonverbal pairs from film/TV scripts and propose to implicitly represent the relation between the verbal and nonverbal messages by situating them in a context. Specifically, we extract *contextualized knowledge* consisting of a verbal statement, its associated nonverbal information, and, as context, the text of the scene in which they occur. To enhance knowledge utilization, we propose a two-stage fine-tuning strategy to use the large-scale weakly-labeled MRC data constructed based on one type of contextualized knowledge and employ a teacher-student paradigm to inject multiple types of contextualized knowledge into a student machine reader. Experimental results show the effectiveness of our method. We present more details in Chapter 6.

1.3 Roadmap

The rest of the dissertation is organized as follows. In Chapter 2, we provide relevant background and related work, covering the history, tasks, and models. In the next four chapters, we tackle previous representative non-extractive MRC tasks (Chapter 3), present new challenges to MRC systems (Chapter 4 and Chapter 5), and present promising directions and approaches for tackling the presented challenges (Chapter 4 and Chapter 6). Finally, in Chapter 7, we summarize the contributions of this dissertation and outline possible future research directions.

CHAPTER 2

BACKGROUND AND RELATED WORK

This chapter presents an overview of the background and existing reading comprehension tasks and models relevant to the work presented in this dissertation. We further discuss related work in greater depth and broader context as needed in the corresponding chapters.

2.1 History

The study of machine reading comprehension dates back to the 1970s when researchers already worked on computer-implemented story comprehension models (Charniak, 1972) and recognized answering questions about paragraphs of text as a task criterion for evaluating language understanding systems' reading skills (Lehnert, 1977). However, the field was mostly neglected in the 1980s and early 1990s (Chen, 2018). Later, the dataset created by Hirschman et al. (1999) spurred a small revival of interest, followed by the ANLP-NAACL 2000 Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems. The dataset is small in scale, and the systems are mostly rule-based. In the 2010s, researchers started to formulate MRC as a supervised learning problem (Richardson et al., 2013; Berant et al., 2014) and create a growing number of large-scale datasets (Rajpurkar et al., 2016; Lai et al., 2017), which greatly facilitate the development of machine learning-based approaches.

It is worth mentioning that machine reading comprehension is closely related

question answering (QA), and nowadays, researchers sometimes use the two terms interchangeably rather than making a clear distinction between them. However, 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 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.

2.2 MRC Tasks

As mentioned in Chapter 1, this dissertation primarily discusses non-extractive multiple-choice MRC, in which answer options are not restricted to extractive text spans. More specifically, we focus on non-extractive multiple-choice MRC tasks that aim to answer ***free-form*** questions, which are not limited to a single question type such as *cloze-style* questions formed by removing a span or a sentence in a text (Hill et al., 2016; Bajgar et al., 2016; Mostafazadeh et al., 2016; Xie et al., 2018; Zheng et al., 2019) or *close-ended* questions that can be answered with a minimal answer (e.g., yes or no (Clark et al., 2019)). It involves extensive human efforts to build such a dataset (e.g., MCTest (Richardson et al., 2013), SemEval-2018 Task 11 (Ostermann et al., 2018), MultiRC (Khashabi et al., 2018), and OpenBookQA (Mihaylov et al., 2018)) by crowdsourcing. Besides crowdsourcing, datasets such as RACE (Lai et al., 2017) and ARC (Clark et al., 2018) are collected from language or science exams designed by educational experts (Penas et al., 2014; Shibuki et al., 2014; Tseng et al., 2016) to evaluate the comprehension

level of human participants. As these kind of datasets are relatively difficult to construct or collect, most existing datasets are small in size, which hinders the development of state-of-the-art deep neural models, compared with large-scale extractive and abstractive MRC datasets (Hermann et al., 2015; Hill et al., 2016; Onishi et al., 2016; Chen and Choi, 2016; Mostafazadeh et al., 2016; Bajgar et al., 2016; Nguyen et al., 2016; Trischler et al., 2017; Joshi et al., 2017; Ma et al., 2018; Kočiskỳ et al., 2018), such as SQuAD (Rajpurkar et al., 2016) and CoQA (Reddy et al., 2019).

2.3 MRC Models

We divide models for MRC into three categories: rule-based model, classical machine learning model, and end-to-end neural model.

Rule-based model. Most early MRC models are based on hand-crafted rules, typically involving rule-based pattern matching (e.g., bag-of-words matching) and shadow linguistic processing (e.g., stemming) (Hirschman et al., 1999; Riloff and Thelen, 2000). For example, typical rule-based models for multiple-choice MRC compute the matching score between each question-option pair and the reference document and choose the option with the highest score as the answer. The matching score is calculated using simple rules such as the count of matched words (Yih et al., 2013) and the sum of the TF-IDF values of the matched words (Richardson et al., 2013).

Classical machine learning model. With the availability of training data, machine learning-based approaches have attracted increasing interest since the 2010s. Early machine learning-based approaches mostly extract rich features and employ classical machine learning algorithms (Sachan et al., 2015; Wang et al., 2015). For example, Wang et al. (2015) develop a model of this type for multiple-choice MRC, which combines features based on rule-based pattern matching, dependency syntax, frame semantics, coreference, and word embeddings in a max-margin learning framework.

End-to-end neural model. The machine learning-based approaches for MRC have been gradually evolving in the direction of deep learning models since 2015, when large-scale training data for MRC started to be available (Hermann et al., 2015). Compared with classical machine learning models, deep learning models rely much less on hand-crafted features. Instead, deep learning models mainly learn the features themselves using end-to-end neural networks. For example, a typical deep learning model for multiple-choice MRC converts the document, question, and option to embedding vectors and passes them to a neural network that consists of several modeling or interaction layers. The neural network is trained to predict if the option is correct (Wang et al., 2018d). Before 2018, the parameters of neural networks in most work are randomly initialized before being tuned using the gradient descent algorithm or its variants on the target MRC task's training data. In 2018, Radford et al. propose first pre-training the neural network with a language model objective over large-scale corpora such as thousands of books and then fine-tuning the pre-trained neural network on the target MRC task. This framework achieves remarkable success in MRC and is generally followed by today's state-of-the-art MRC models.

CHAPTER 3

READING STRATEGIES FOR IMPROVED READING COMPREHENSION

Recently, significant progress has been achieved on many natural language processing tasks including MRC by fine-tuning a pre-trained general-purpose language model (Radford et al., 2018; Devlin et al., 2019). However, similar to the process of knowledge accumulation for human readers, it is time-consuming and resource-demanding to impart massive amounts of general domain knowledge from external corpora into a deep language model via pre-training. For example, it takes a month to pre-train a 12-layer transformer on eight P100 GPUs over the BooksCorpus (Zhu et al., 2015; Radford et al., 2018); Devlin et al. (2019) pre-train a 24-layer transformer using 64 TPUs for four days on the BooksCorpus plus English Wikipedia, a feat not easily reproducible considering the tremendous computational resources (≈ one year to train on eight P100 GPUs).

From a practical viewpoint, given a limited number of training instances and a pre-trained model, can we improve machine reading comprehension during fine-tuning instead of imparting more prior knowledge into a model via expensive pre-training? Inspired by reading strategies identified in cognitive science research that have been shown effective in improving comprehension levels of human readers, especially those who lack adequate prior knowledge of the topic of the text (Mokhtari and Sheorey, 2002; Mokhtari and Reichard, 2002; McNamara et al., 2004), we propose in this chapter three corresponding domain-independent strategies to improve MRC based on an existing pre-trained transformer (Section 3.2.1):

- BACK AND FORTH READING (*"I go back and forth in the text to find relationships among ideas in it."*):

consider both the original and reverse order of an input sequence (Section 3.2.2)

- HIGHLIGHTING (*"I highlight information in the text to help me remember it."*): add a trainable embedding to the text embedding of those tokens deemed relevant to the question and candidate answers (Section 3.2.3)
- SELF-ASSESSMENT (*"I ask myself questions I would like to have answered in the text, and then I check to see if my guesses about the text are right or wrong."*): generate practice questions and their associated span-based candidate answers from the existing reference documents (Section 3.2.4)

By fine-tuning a pre-trained transformer (Radford et al., 2018) according to our proposed strategies on the largest general domain multiple-choice MRC dataset RACE (Lai et al., 2017) collected from language exams, we obtain a 5.8% absolute improvement in accuracy over the previous best result achieved by the same pre-trained transformer fine-tuned on RACE without the use of strategies (Section 3.3.2). We further fine-tune the resulting model on a target MRC task. Experiments show that our method achieves new state-of-the-art results on six representative non-extractive MRC datasets that require a range of reading skills such as commonsense and multi-sentence reasoning (i.e., ARC (Clark et al., 2016, 2018), OpenBookQA (Mihaylov et al., 2018), MCTest (Richardson et al., 2013), SemEval-2018 Task 11 (Yang et al., 2017), ROCStories (Mostafazadeh et al., 2016), and MultiRC (Khashabi et al., 2018)) (Section 3.3.4). These results indicate the effectiveness of our proposed strategies and the versatility and generality of our fine-tuned models that incorporate the strategies.

This chapter is based on Sun et al. (2019b).

3.1 Task Introduction

In this chapter, we investigate how to make use of limited resources to improve MRC, using seven representative multiple-choice MRC datasets as case studies. As shown in Table 3.1, the majority of the correct answer options in most of the datasets (except for ARC and MCTest) are non-extractive. Except for MultiRC, there is exactly one correct answer option for each question. For ARC and OpenBookQA, a reference corpus is provided instead of a single reference document associated with each question.

Here we give a formal **task definition**. Given a reference document d, a question q, and associated answer options $\{o_1, o_2, \ldots, o_m\}$, the goal is to select the correct answer option(s). We can easily adapt our method to an MRC task that only provides a reference corpus (Section 3.3.4).

3.2 Approach

We first introduce a neural reader based on a pre-trained transformer (Section 3.2.1) and then elaborate on the strategies that are applied during the fine-tuning stage — back and forth reading (Section 3.2.2), highlighting (Section 3.2.3), and self-assessment (Section 3.2.4).

3.2.1 Framework Overview

Our neural reader follows the framework of discriminatively fine-tuning a generative pre-trained transformer (GPT) (Radford et al., 2018). It adapts a pre-trained

	RACE	**ARC**	**OpenBookQA**	**MCTest**	**SemEval-2018 Task 11**	**ROCStories**	**MultiRC**
construction method	exams	exams	crowd.	crowd.	crowd.	crowd.	crowd.
sources of documents	general	science	science	stories	narrative text	stories	mixed-domain
average # of answer options	4.0	4.0	4.0	4.0	2.0	2.0	5.4
# of documents	27,933	14M†	1,326†	660	2,119	3,742	871
# of questions	**97,687**	7,787	5,957	2,640	13,939	–	9,872
non-extractive answer⋆ (%)	87.0	43.3	83.8	45.3	89.9	100.0	82.1

Table 3.1: Statistics of multiple-choice machine reading comprehension datasets. Some values come from Reddy et al. (2019), Kočiskỳ et al. (2018), and Lai et al. (2017) (crowd.: crowdsourcing; †: regarding each sentence/claim as a document (Clark et al., 2018); ⋆: correct answer options that are not text snippets from reference documents).

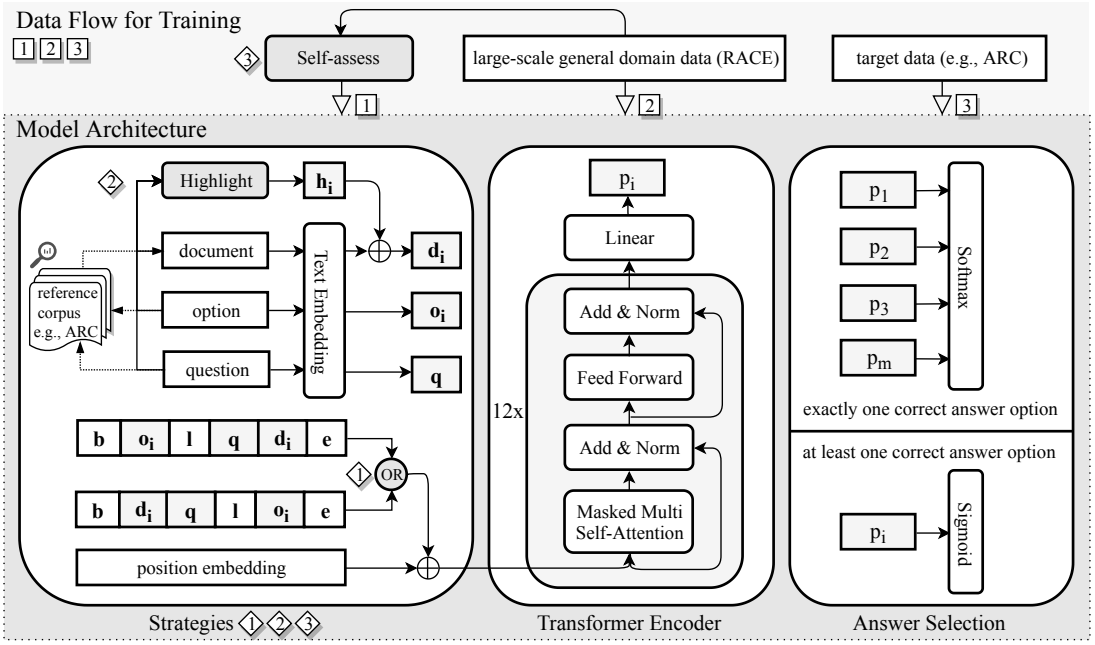


Figure 3.1: Framework Overview. Strategy 1, 2, and 3 refer to back and forth reading (BF) (Section 3.2.2), highlighting (HL) (Section 3.2.3), and self-assessment (SA) (Section 3.2.4), respectively.

multi-layer transformer (Vaswani et al., 2017; Liu et al., 2018) language model to a labeled dataset C, where each instance consists of a sequence of input tokens $x^1, \ldots, x^n$, along with a label y, by maximizing:

$$\sum_{x,y} \log P(y \mid x^1, \ldots, x^n) + \lambda \cdot L(C) \tag{3.1}$$

where L is the likelihood from the language model, λ is the weight of language model, and $P(y \mid x^1, \ldots, x^n)$ is obtained by a linear classification layer over the final transformer block's activation of the language model. For multiple-choice MRC tasks, $x^1, \ldots, x^n$ come from the concatenation of a start token, a reference document, a question, a delimiter token, an answer option, and an end token; y indicates the correctness of an answer option. We refer readers to Radford et al. (2018) for more details.

Apart from placing a delimiter to separate the answer option from the docu-

ment and question, the original framework pays little attention to task-specific structures in MRC tasks. Inspired by reading strategies, with limited resources and a pre-trained transformer, we propose three strategies to improve machine reading comprehension. We show the whole framework in Figure 3.1.

3.2.2 Back and Forth Reading (BF)

For simplicity, we represent the original input sequence of GPT during fine-tuning (Radford et al., 2018) as [$dq \$ o$], where [, $, and] represent the start token, delimiter token, and end token, respectively. Inspired by back and forth reading, we consider both the original order and the reverse order [$o \$ qd$]. The token order within d, q, and o is still preserved. We fine-tune two GPTs that use [$dq \$ o$] and [$o \$ qd$] as the input sequence respectively, and then we ensemble the two models. We also consider other similar pairs of input sequences such as [$qd \$ o$] and [$o \$ dq$] in the experiments (Section 3.3.3).

3.2.3 Highlighting (HL)

In the original implementation (Radford et al., 2018), during the fine-tuning stage of GPT, the text embedding of a document is independent of its associated questions and answer options. Inspired by highlights used in human reading, we aim to make the document encoding aware of the associated question-answer option pair (q, o_i). We focus on the content words in questions and answer options since they appear to provide more useful information (Mirza and Bernardi, 2013), and we identify them via their part of speech (POS) tags, one of: noun, verb,

adjective, adverb, numeral, or foreign word.

Formally, we let T be the set of POS tags of the content words. We let $\boldsymbol{d}$ denote the sequence of the text embedding of document d. We use d^j to represent the j^{th} token in d and $\boldsymbol{d^j}$ to denote the text embedding of d^j. Given d and a (q, o_i) pair, we define a *highlight embedding* $\boldsymbol{h_i^j}$ for the j^{th} token in d as:

$$\boldsymbol{h_i^j} = \begin{cases} \boldsymbol{\ell^+} & \text{if the POS tag of } d^j \text{ belongs to } T, \\ & \text{and } d^j \text{ appears in either } q \text{ or } o_i \\ \boldsymbol{\ell^-} & \text{otherwise} \end{cases} \tag{3.2}$$

where $\boldsymbol{\ell^+}$ and $\boldsymbol{\ell^-}$ are two trainable vectors of the same dimension as $\boldsymbol{d^j}$.

Following the above definition, the sequence of the highlight embedding $\boldsymbol{h_i} = \boldsymbol{h_i^1}, \boldsymbol{h_i^2}, \ldots, \boldsymbol{h_i^n}$ is of the same length as $\boldsymbol{d}$. We replace $\boldsymbol{d}$ with $\boldsymbol{d_i} = \boldsymbol{d} + \boldsymbol{h_i}$ when we encode a document. More specifically, we use the concatenation of $\boldsymbol{b}$, $\boldsymbol{d_i}$, $\boldsymbol{q}$, $\boldsymbol{l}$, $\boldsymbol{o_i}$, and $\boldsymbol{e}$ as the new input of GPT during fine-tuning (Section 3.2.1), where $\boldsymbol{b}$, $\boldsymbol{l}$, and $\boldsymbol{e}$ denote the embedding of the start token, delimiter token, and end token, respectively, and $\boldsymbol{q}$ and $\boldsymbol{o_i}$ represent the sequence of the text embedding of q and o_i, respectively.

3.2.4 Self-Assessment (SA)

While in previous work (Radford et al., 2018), the original GPT is directly fine-tuned on an MRC end task, we instead develop a fine-tuning approach inspired by the self-assessment reading strategy. In particular, we propose a simple method to generate questions and their associated multiple span-based answer options, which cover the content of multiple sentences from a reference document.

By first fine-tuning a pre-trained model on these *practice* instances, we aim to render the resulting fine-tuned model more aware of the input structure and to integrate information across multiple sentences as may be required to answer a given question.

Concretely, we randomly generate no more than n_q questions and associated answer options based on each document from the end task (i.e., RACE in this chapter). We describe the steps as follows.

- **Input:** a reference document from the end task.
- **Output:** a question and four answer options associated with the reference document.

1. Randomly pick no more than n_s sentences from the document and concatenate these sentences together.
2. Randomly pick no more than n_c non-overlapping spans from the concatenated sentences. Each span randomly contains no more than n_t tokens within a single sentence. We concatenate the selected spans to form the correct answer option. We remove the selected spans from the concatenated sentences and use the remaining text as the question.
3. Generate three distractors (i.e., wrong answer options) by randomly replacing spans in the correct answer option with randomly picked spans from the document.

where n_q, n_s, n_c, and n_t are used to control the number and difficulty level of the questions.

3.3 Experiment

3.3.1 Experiment Settings

For most of the hyperparameters, we follow the work of Radford et al. (2018). We use the same preprocessing procedure and the released pre-trained transformer. We generate 119k instances based on the reference documents from the training and development set of RACE (Lai et al., 2017), with $n_q = 10$, $n_s = 3$, $n_c = 4$, and $n_t = 4$ (Section 3.2.4). We first fine-tune the original pre-trained model on these automatically generated instances with 1 training epoch (data flow 1 boxed in Figure 3.1). We then fine-tune the model on a large-scale general domain MRC dataset RACE with 5 training epochs (data flow 2 boxed in Figure 3.1). Finally, we fine-tune the resulting model on one of the aforementioned six out-of-domain MRC datasets (at max 10 epochs). See data flow 3 boxed in Figure 3.1. When we fine-tune the model on different datasets, we set the batch size to 8, language model weight λ to 2. We ensemble models by averaging logits after the linear layer. For strategy highlighting (Section 3.2.3), the content-word POS tagset T = {NN, NNP, NNPS, NNS, VB, VBD, VBG, VBN, VBP, VBZ, JJ, JJR, JJS, RB, RBR, RBS, CD, FW}, and we randomly initialize $\boldsymbol{\ell}^+$ and $\boldsymbol{\ell}^-$.

3.3.2 Evaluation on RACE

In Table 3.2, we first report the accuracy of the state-of-the-art models (MMN and original fine-tuned GPT) and Amazon Turkers (Human performance). We then report the performance of our implemented fine-tuned GPT baselines and our models (GPT+Strategies). Results are shown on the RACE dataset (Lai et al.,

Approach		#	RACE-M	RACE-H	RACE
MMN (Tang et al., 2019)		9	64.7	55.5	58.2
GPT (Radford et al., 2018)		1	62.9	57.4	59.0
Human performance (Lai et al., 2017)		1	85.1	69.4	73.3
GPT⋆		1	60.9	57.8	58.7
		2	62.6	58.4	59.6
		9	63.5	59.3	60.6
GPT⋆ + Strategies	SA	1	63.2	59.2	60.4
	HL	1	67.4	61.5	63.2
	BF	2	67.3	60.7	62.6
	SA + HL	1	**69.2**	**61.5**	**63.8**
	SA + HL + BF	2	**70.9**	**63.2**	**65.4**
	SA + HL + BF	9	**72.0**	**64.5**	**66.7**

Table 3.2: Accuracy (%) on the test set of RACE (#: number of (ensemble) models; SA: Self-Assessment; HL: Highlighting; BF: Back and Forth Reading; ⋆: our implementation).

2017) and its two subtasks: RACE-M collected from middle school exams and RACE-H collected from high school exams.

Our single and ensemble models outperform previous state-of-the-art (i.e., GPT and GPT (9×)) by a large margin (63.8% vs. 59.0%; 66.7% vs. 60.6%). The two single-model strategies – self-assessment and highlighting – improve over the single-model fine-tuned GPT baseline (58.7%) by 1.7% and 4.5%, respectively. Using the back and forth reading strategy, which involves two models, gives a 3.0% improvement in accuracy compared to the ensemble of two original fine-tuned GPTs (59.6%). Strategy combination further boosts the performance. By combining self-assessment and highlighting, our single model achieves a 5.1% improvement in accuracy over the fine-tuned GPT baseline (63.8% vs. 58.7%). We apply all the strategies by ensembling two such single models that read an input sequence in either the original or the reverse order, leading to a 5.8% improvement in accuracy over the ensemble of two original fine-tuned GPTs (65.4% vs. 59.6%).

To further analyze performance, we roughly divide the question types into five categories: detail (*facts and details*), inference (*reasoning ability*), main (*main idea or purpose of a document*), attitude (*author's attitude toward a topic or tone/source of a document*), and vocabulary (*vocabulary questions*) (Qian and Schedl, 2004; Lai et al., 2017) and annotate all the instances of the RACE development set. As shown in Figure 3.2, compared to the fine-tuned GPT baseline, our single-model strategies (SA and HL) consistently improve the results across all categories. Compared to other strategies, highlighting is likely to lead to bigger gains for most question types.

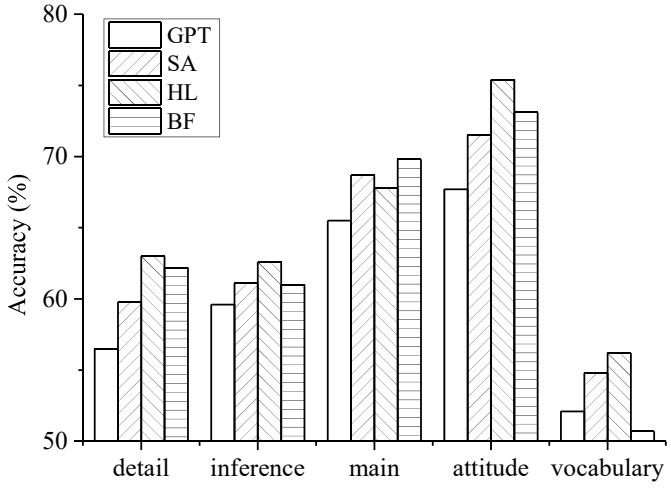


Figure 3.2: Performance on different question types.

Compared to human performance, there is still a considerable room for improvements, especially on RACE-M. We take a close look at the instances from the RACE-M development set that all our implementations fail to answer correctly. We notice that 82.0% of them require one or multiple types of world knowledge (e.g., negation resolution, commonsense, paraphrase, and mathematical/logic knowledge (Sugawara et al., 2017b,a, 2018)), especially when correct answer options are not explicitly mentioned in the reference document. For example, we need the knowledge — *the type of thing that is written by a writer can probably*

be a book — to answer the question *"follow your heart is a _"* from the context *"Follow your heart by Andrew Matthews, an Australian writer, tells us that making our dreams real is life's biggest challenge"*. Besides, 19.7% of these failed instances require coreference resolution. It might be promising to leverage coreference resolvers to connect nonadjacent relevant sentences.

Task	Metric	Previous STOA		GPT	GPT (2×)	GPT +Strategies	GPT +Strategies (2×)
ARC-Easy	Acc.	Clark et al. (2018)	62.6	57.0	57.1	66.6	68.9
ARC-Challenge	Acc.	Ni et al. (2018)	36.6	38.2	38.4	40.7	42.3
OpenBookQA	Acc.	Mihaylov et al. (2018)	50.2	52.0	52.8	55.2	55.8
MCTest-MC160	Acc.	Chung et al. (2018)	76.4	65.4	65.8	80.0	81.7
MCTest-MC500	Acc.	Chung et al. (2018)	72.3	61.5	61.0	78.7	82.0
SemEval	Acc.	Chen et al. (2018)	84.1	88.0	88.6	88.8	89.5
ROCStories	Acc.	Radford et al. (2018)	86.5	87.1	87.5	88.0	88.3
MultiRC	$F1_m$	Khashabi et al. (2018)	66.5	69.3	70.3	71.5	73.1
	$F1_a$	Khashabi et al. (2018)	63.2	67.2	67.7	69.2	70.5
	Acc.†	Khashabi et al. (2018)	11.8	15.2	16.5	22.6	21.8
Average	Acc.		60.1	58.1	58.5	**65.1**	**66.3**

Table 3.3: Performance (%) on the test sets of ARC, OpenBookQA, MCTest, SemEval-2018 Task 11, and ROCStories and the development set of MultiRC (Acc.: Accuracy; $F1_m$: macro-average F1; $F1_a$: micro-average F1; †: using the joint exact match accuracy (i.e., EM_0 reported by the official evaluation (Khashabi et al., 2018))). RACE is used as the source task for all our implementations.

3.3.3 Further Discussions on Strategies

Besides the strategies introduced in Section 3.2, we also explore other reading strategies such as SUMMARIZATION (*"I take an overall view of the text to see what it is about before carefully reading it."*) by appending an extractive summary (Boudin et al., 2015) before each reference document, which is shown less effective for machine reading comprehension in our experiments compared to the strategies we focus on. In this section, we further discuss the three strategies.

Back and Forth Reading We notice that the input order difference between

Task	Metric	GPT	GPT (2×)	GPT +Strategies	GPT +Strategies (2×)
ARC-Easy	Acc.	54.0	53.9	61.9	63.1
ARC-Challenge	Acc.	30.3	30.7	35.0	35.4
OpenBookQA	Acc.	50.0	50.0	54.2	55.0
MCTest-MC160	Acc.	58.8	60.0	67.5	70.8
MCTest-MC500	Acc.	52.0	54.0	64.7	64.8
SemEval	Acc.	87.3	88.0	87.6	88.1
ROCStories	Acc.	86.7	87.0	87.4	88.1
MultiRC	$F1_m$	69.3	69.3	68.8	69.7
	$F1_a$	66.2	66.5	67.4	67.9
	Acc.†	11.9	13.1	16.2	16.9
Average	Acc.	53.9	54.6	**59.3**	**60.3**

Table 3.4: Performance (%) on the test sets of ARC, OpenBookQA, MCTest, SemEval-2018 Task 11, and ROCStories and the development set of MultiRC using the target data only (i.e., without the data flow 1 and 2 boxed in Figure 3.1) (Acc.: Accuracy; $F1_m$: macro-average F1; $F1_a$: micro-average F1; †: using the joint exact match accuracy (i.e., EM_0 reported by the official evaluation (Khashabi et al., 2018))).

two ensemble models is likely to yield performance gains. Besides ensembling two models that use input sequence [$dq \,\$\, o$] and [$o \,\$\, qd$] respectively, we also investigate other reverse or almost reverse pairs. For example, we can achieve better results by ensembling [$qd \,\$\, o$] and [$o \,\$\, dq$] (61.0%) or [$qd \,\$\, o$] and [$o \,\$\, qd$] (61.7%), compared to the ensemble of two original fine-tuned GPTs (both of them use [$d \,\$\, qo$]) on the RACE dataset (59.6% in Table 3.2).

Highlighting We try two variants to define highlight embeddings (Equation 3.2 in Section 3.2.3) by considering the content of questions only or answer options only. Experiments show that using partial information yields a decrease in accuracy (60.6% and 61.0%, respectively) compared to 63.2% (Table 3.2), achieved by considering the content words in a question and its answer options. We attempt to also highlight the coreferential mentions of the content words, which does not lead to further gains, though.

Self-Assessment We explore alternative approaches to generate questions.

For example, we use the Wikipedia articles from SQuAD (Rajpurkar et al., 2016) instead of the general domain documents from the end task RACE. We generate the same number of questions as the number of questions we generate using RACE following the same steps mentioned in Section 3.2.4. Experiments show that this method also improves the accuracy of the fine-tuned GPT baseline (59.7% vs. 58.7%). As self-assessment can be somehow regarded as a data augmentation method, we investigate other unsupervised question generation methods such as sentence shuffling and paraphrasing via back-translation (Ding and Zhou, 2018; Yu et al., 2018). Our experiments demonstrate that neither of them results in performance improvements on the RACE dataset.

3.3.4 Adaptation to Other Non-Extractive MRC Tasks

We follow the philosophy of transferring the knowledge from a high-performing model pre-trained on a large-scale supervised data of a source task to a target task, in which only a small amount of training data is available (Chung et al., 2018). RACE has been used to pre-train a model for other MRC tasks as it contains the largest number of general domain non-extractive questions (Table 3.1) (Ostermann et al., 2018; Wang et al., 2018c). In our experiment, we also treat RACE as the source task and regard six representative non-extractive multiple-choice MRC datasets from multiple domains as the target tasks.

We require some task-specific modifications considering the different structures of these datasets. In ARC and OpenBookQA, there is no reference document associated with each question. Instead, a reference corpus is provided, which consists of unordered science-related sentences relevant to questions. We there-

fore first use Lucene (McCandless et al., 2010) to retrieve the top 50 sentences by using the non-stop words in a question and one of its answer options as a query. The retrieved sentences are used to form the reference document for each answer option. In MultiRC, a question could have more than one correct answer option. Therefore, we use a sigmoid function instead of softmax at the final layer (Figure 3.1) and regard the task as a binary (i.e., correct or incorrect) classification problem over each (document, question, answer option) instance. When we adapt our method to the non-conventional MRC dataset ROCStories, which aims at choosing the correct ending to a four-sentence incomplete story from two answer options (Mostafazadeh et al., 2016), we leave the question context empty as no explicit questions are provided. Since the test set of MultiRC is not publicly available, we report the performance of the model that achieves the highest micro-average F1 ($F1_a$) on the development set. For other tasks, we select the model that achieves the highest accuracy on the development set and report the accuracy on the test set.

We first fine-tune GPT using our proposed three strategies on RACE and further fine-tune the resulting model on one of the six target tasks (see Table 3.3). During the latter fine-tuning stage, besides the *highlighting* embeddings inherited from the previous fine-tuning stage, we also apply the strategy *back and forth reading*, and we do not consider *self-assessment* since the model has already benefited from the high-quality RACE instances during the first fine-tuning stage. We compare with the baselines that are first fine-tuned on RACE and then fine-tuned on a target task without the use of strategies, which already outperform previous state-of-the-art (SOTA) on four out of six datasets (OpenBookQA, SemEval-2018 Task 11, ROCStories, and MultiRC). By using the strategies, we obtain a 7.8% absolute improvement in average accuracy over the ensemble baseline (58.5%)

and a 6.2% absolute improvement over previous SOTA (60.1%).

To further investigate the contribution of the strategies, we directly fine-tune GPT on a target task without using the labeled data in RACE (i.e., we only keep data flow 3 in Figure 3.1). Compared to the baseline that is fine-tuned without using strategies (54.6%), we obtain a 10.4% relative improvement in average accuracy (60.3%) and especially large improvements on datasets ARC, OpenBookQA, and MCTest (Table 3.4).

3.4 Related Work

3.4.1 Methods for Multiple-Choice MRC

We primarily discuss methods applied to large-scale datasets such as RACE (Lai et al., 2017). Researchers develop a variety of methods with attention mechanisms (Chen et al., 2016; Dhingra et al., 2017; Xu et al., 2018b; Tay et al., 2018; Tang et al., 2019) for improvement such as adding an elimination module (Parikh et al., 2018) or applying hierarchical attention strategies (Zhu et al., 2018; Wang et al., 2018d). These methods seldom take the rich external knowledge (other than pre-trained word embeddings) into considerations. Instead, we investigate different strategies based on an existing pre-trained transformer (Radford et al., 2018) (Section 3.2.1), which leverages rich linguistic knowledge from external corpora and achieves state-of-the-art performance on a wide range of natural language processing tasks including machine reading comprehension.

3.4.2 Transfer Learning for MRC and QA

Transfer learning techniques have been successfully applied to machine reading comprehension (Golub et al., 2017; Chung et al., 2018) and question answering (Min et al., 2017; Wiese et al., 2017). Compared to previous work, we simply fine-tune our model on the source data and then further fine-tune the entire model on the target data. The investigation of methods such as adding additional parameters or an L2 loss and fine-tuning only part of the parameters is beyond the scope of this work.

3.4.3 Data Augmentation for MRC Without Using External Datasets

Previous methods augment the training data for extractive machine reading comprehension and question answering by randomly reordering words or shuffling sentences (Ding and Zhou, 2018; Li and Zhou, 2018) or generating questions through paraphrasing (Yang et al., 2017; Yuan et al., 2017), which require a large amount of training data or limited by the number of training instances (Yu et al., 2018). In comparison, our problem (i.e., question and answer options) generation method does not rely on any existing questions in the training set, and the generated questions can involve the content of multiple sentences in a reference document.

3.5 Chapter Summary

Inspired by previous research on reading strategies for improved comprehension levels of human readers, we propose three strategies (i.e., back and forth reading, highlighting, and self-assessment), aiming at improving machine reading comprehension using limited resources: a pre-trained language model and a limited number of training instances. By applying the proposed three strategies, we obtain a 5.8% absolute improvement in accuracy over the state-of-the-art performance on the RACE dataset. By fine-tuning the resulting model on a new target task, we achieve new state-of-the-art results on six representative non-extractive MRC datasets from multiple domains that require a diverse range of reading skills. These results consistently indicate the effectiveness of our proposed strategies and the general applicability of our fine-tuned model that incorporates these strategies.

CHAPTER 4

DIALOGUE-BASED READING COMPREHENSION

In the preceding chapter, our study is carried out on previous representative non-extractive MRC datasets. Source documents in these datasets have generally been drawn from formal written texts such as news, fiction, and Wikipedia articles, which are commonly considered well-written, accurate, and neutral (Lai et al., 2017; Khashabi et al., 2018; Ostermann et al., 2018).

With the goal of advancing research in machine reading comprehension and facilitating dialogue understanding, we construct and present in this chapter DREAM — the first multiple-choice **D**ialogue-based **REA**ding comprehension exa**M**ination dataset. We collect 10,197 questions for 6,444 multi-turn multi-party dialogues from English language exams, which are carefully designed by educational experts (e.g., English teachers) to assess the comprehension level of Chinese learners of English. Each question is associated with three answer options, exactly one of which is correct. (See Table 4.1 for an example.) DREAM covers a variety of topics and scenarios in daily life such as conversations on the street, on the phone, in a classroom or library, at the airport or the office or a shop (Section 4.1).

Based on our analysis of DREAM, we argue that dialogue-based reading comprehension is at least as difficult as existing non-conversational counterparts. In particular, answering **34%** of DREAM questions requires unspoken commonsense knowledge, e.g., unspoken scene information. This might be due to the nature of dialogues: for efficient oral communication, people rarely state obvious explicit world knowledge (Forbes and Choi, 2017) such as *"Christmas Day is celebrated on December 25th"*. Understanding the social implications of an

Dialogue 1 (D1)	
W:	Tom, look at your shoes. How dirty they are! You must clean them.
M:	Oh, mum, I just cleaned them yesterday.
W:	They are dirty now. You must clean them again.
M:	I do not want to clean them today. Even if I clean them today, they will get dirty again tomorrow.
W:	All right, then.
M:	Mum, give me something to eat, please.
W:	You had your breakfast in the morning, Tom, and you had lunch at school.
M:	I am hungry again.
W:	Oh, hungry? But if I give you something to eat today, you will be hungry again tomorrow.
Q1	Why did the woman say that she wouldn't give him anything to eat?
A.	Because his mother wants to correct his bad habit.⋆
B.	Because he had lunch at school.
C.	Because his mother wants to leave him hungry.

Table 4.1: A sample DREAM problem that requires general world knowledge (⋆: the correct answer option).

utterance as well as inferring a speaker's intentions is also regularly required for answering dialogue-based questions. The dialogue content in Table 4.1, for example, is itself insufficient for readers to recognize the intention of the female speaker (W) in the first question (Q1). However, world knowledge is rarely considered in state-of-the-art reading comprehension models (Tay et al., 2018; Wang et al., 2018d).

Moreover, dialogue-based questions can cover information imparted across multiple turns involving multiple speakers. In DREAM, approximately **85%** of questions can only be answered by considering the information from multiple sentences. For example, to answer Q1 in Table 4.2 regarding the date of birth of the male speaker (M), the supporting sentences (in bold) include *"You know, tomorrow is Christmas Day"* from the female speaker and *"...I am more than excited about my birthday, which will come in two days"* from the male speaker. Compared to "multiple-sentence questions" in traditional reading comprehension datasets, DREAM further requires an understanding of the turn-based structure of dialogue, e.g. for aligning utterances with their corresponding speakers.

As only **16%** of correct answer options are text spans from the source documents, we primarily explore rule-based methods and state-of-the-art neural models designed for multiple-choice reading comprehension (Section 4.2). We find first that neural models designed for non-dialogue-based reading comprehension (Chen et al., 2016; Dhingra et al., 2017; Wang et al., 2018d) do not fare well: the highest achieved accuracy is 45.5%, only slightly better than the accuracy (44.6%) of a simple lexical baseline (Richardson et al., 2013). For the most part, these models fundamentally exploit only surface-level information from the source documents. Considering the above-mentioned challenges, however, we hypothesize that incorporating general world knowledge and aspects of the dialogue structure would allow a better understanding of the dialogues. As a result, we modify our baseline systems to include (1) general world knowledge in the form of such as ConceptNet relations (Speer et al., 2017) and a pre-trained language model (Radford et al., 2018), and (2) speaker information for each utterance. Experiments show the effectiveness of these factors on the lexical baselines as well as neural and non-neural machine learning approaches: we acquire up to 11.9% absolute gain in accuracy compared to the highest performance achieved by the state-of-the-art reading comprehension model (Wang et al., 2018d) that mainly relies on explicit surface-level information in the text (Section 4.3).

Finally, we see a significant gap between the best automated approach (59.5%) and human ceiling performance (98.6%) on the DREAM dataset. This provides yet additional evidence that dialogue-based reading comprehension is a very challenging task. We hope that it also inspires the research community to develop methods for the dialogue-based reading comprehension task.

This chapter is based on Sun et al. (2019a).

4.1 Data

In this section, we describe how we construct DREAM (Section 4.1.1) and provide a detailed analysis of this dataset (Section 4.1.2).

Dialogue 2 (D2)	
W:	Hey, Mike. Where have you been? I didn't see you around these days?
M:	I was hiding in my office. My boss gave me loads of work to do, and I tried to finish it before my birthday. Anyway, I am done now. Thank goodness! How is everything going with you?
W:	I'm quite well. **You know, tomorrow is Christmas Day.** Do you have any plans?
M:	**Well, to tell you the truth, I am more than excited about my birthday, which will come in two days.** I am going to visit my parents-in-law with my wife.
W:	Wow, sounds great.
M:	Definitely! This is my first time to spend my birthday with them.
W:	Do they live far away from here?
M:	A little bit. We planned to take the train, but considering the travel peak, my wife strongly suggested that we go to the airport right after we finish our work this afternoon. How about you? What's your holiday plan?
W:	Well, our situations are just the opposite. My parents-in-law will come to my house, and they wish to stay at home and have a quiet Christmas Day. So I have to call my friends to cancel our party that will be held at my house.
M:	You'll experience a quite different and lovely holiday. Enjoy your Christmas!
W:	Thanks, the same to you!
Q1	What is the date of the man's birthday?
A.	25th, December.
B.	26th, December.⋆
C.	27th, December.
Q2	How will the man go to his wife's parents' home?
A.	By train.
B.	By bus.
C.	By plane.⋆
Q3	What is the probable relationship between the two speakers?
A.	Husband and wife.
B.	Friends.⋆
C.	Parent-in-law and son-in-law.

Table 4.2: A complete sample DREAM problem (⋆: the correct answer option).

4.1.1 Collection Methodology

We collect dialogue-based comprehension problems from a variety of English language exams (including practice exams) such as National College Entrance

Examination, College English Test, and Public English Test[1], which are designed by human experts to assess either the listening or reading comprehension level of Chinese English learners in high schools and colleges (aged 12-22). All the problems in DREAM are freely accessible online for public usage. Each problem consists of a dialogue and a series of multiple-choice questions. To ensure every question is associated with exactly three answer options, we drop wrong answer option(s) randomly for questions with more than three options. We remove duplicate problems and randomly split the data at the problem level, with 60% train, 20% development, and 20% test.

4.1.2 Data Analysis

We summarize the statistics of DREAM in Table 4.3 and data split in Table 4.4. Compared to existing datasets built from formal written texts, the vocabulary size is relatively small since spoken English by its nature makes greater use of high-frequency words and needs a smaller vocabulary for efficient real-time communication (Nation, 2006).

Metric	**Value**
# of answer options per question	3
# of turns	30,183
Avg./Max. # of questions per dialogue	1.6 / 10
Avg./Max. # of speakers per dialogue	2.0 / 7
Avg./Max. # of turns per dialogue	4.7 / 48
Avg./Max. option length (in tokens)	5.3 / 21
Avg./Max. question length (in tokens)	8.6 / 24
Avg./Max. dialogue length (in tokens)	85.9 / 1,290
vocabulary size	13,037

Table 4.3: The overall statistics of DREAM. A turn is defined as an uninterrupted stream of speech from one speaker in a dialogue.

[1]We list all the websites used for data collection in the released dataset.

	Train	Dev	Test	All
# of dialogues	3,869	1,288	1,287	6,444
# of questions	6,116	2,040	2,041	10,197

Table 4.4: The separation of the training, development, and test sets in DREAM.

We categorize questions into two main categories according to the types of knowledge required to answer them: *matching* and *reasoning*.

- **Matching** A question is entailed or paraphrased by exactly one sentence in a dialogue. The answer can be extracted from the same sentence. For example, we can easily verify the correctness of the question-answer pair (*"What kind of room does the man want to rent?"*, *"A two-bedroom apartment."*) based on the sentence *"**M**: I'm interested in renting a two-bedroom apartment"*. This category is further divided into two categories *word matching* and *paraphrasing* in previous work (Chen et al., 2016; Trischler et al., 2017).
- **Reasoning** Questions that cannot be answered by the surface meaning of a single sentence belong to this category. We further define four subcategories as follows.
  - **Summary** Answering this kind of questions requires the whole picture of a dialogue, such as the topic of a dialogue and the relation between speakers (e.g., D2-Q3 in Table 4.2). Under this category, questions such as *"What are the two speakers talking about?"* and *"What are the speakers probably doing?"* are frequently asked.
  - **Logic** We require logical reasoning to answer questions in this category. We usually need to identify logically implied relations among multiple sentences in a dialogue. To reduce the ambiguity during the annotation, we regard a question that can only be solved by considering the content

of multiple sentences and does not belong to the *summary* subcategory that involves all the sentences in a dialogue as a *logic* question. Following this definition, both D2-Q1 and D2-Q2 in Table 4.2 belong to this category.

- **Arithmetic** Inferring the answer requires arithmetic knowledge (e.g., D2-Q1 in Table 4.2 requires $25 - 1 + 2 = 26$).
- **Commonsense** To answer questions under this subcategory, besides the textual information in the dialogue, we also require external commonsense knowledge that cannot be obtained from the dialogue. For instance, all questions in Table 4.2 fall under this category. D2-Q1 and D2-Q2 in Table 4.2 belong to both *logic* and *commonsense* since they require multiple sentences as well as commonsense knowledge for question answering. There exist multiple types of commonsense knowledge in DREAM such as the well-known properties of a highly-recognizable entity (e.g., D2-Q1 in Table 4.2), the prominent relationship between two speakers (e.g., D2-Q3 in Table 4.2), the knowledge of or shared by a particular culture (e.g., when a speaker says *"Cola? I think it tastes like medicine."*, she/he probably means *"I don't like cola."*), and the cause-effect relation between events (e.g., D1-Q1 in Table 4.1). We refer readers to LoBue and Yates (2011) for detailed definitions.

Table 4.5 shows the question type distribution labeled by two human annotators on 25% questions randomly sampled from the development and test sets. Besides the previously defined question categories, we also report the percentage of questions that require reasoning over multiple sentences (i.e., *summary* or *logic* questions) and the percentage of questions that require the surface-level understanding or commonsense/math knowledge based on the content of a single sentence. As a question can belong to multiple reasoning subcategories, the

summation of the percentage of reasoning subcategories is not equal to the percentage of reasoning. The Cohen's kappa coefficient is 0.67 on the development set and 0.68 on the test set.

Question Type	Dev	Test	Dev + Test
Matching	13.0	10.3	11.7
Reasoning	87.0	89.7	88.3
Summary	8.4	15.9	12.1
Logic	74.5	70.4	72.5
Arithmetic	5.1	3.6	4.4
Commonsense	31.5	35.9	33.7
Single sentence	17.1	13.7	15.4
Multiple sentences	82.9	86.3	84.6

Table 4.5: Distribution (%) of question types.

Dialogues in DREAM are generally clean and mostly error-free since they are carefully designed by educational experts. However, it is not guaranteed that each dialogue is written or proofread by a native speaker. Besides, dialogues tend to be more proper and less informal for exam purposes. To have a rough estimation of the quality of dialogues in DREAM and the differences between these dialogues and more casual ones in movies or TV shows, we run a proofreading tool – Grammarly[2] – on all the dialogues from the annotated 25% instances of the development set and the same size (20.7k tokens) of dialogues from *Friends*, a famous American TV show whose transcripts are commonly used for dialogue understanding (Chen and Choi, 2016; Ma et al., 2018). As shown in Table 4.6, there exist fewer spelling mistakes and the overall score is slightly higher than that of the dialogues in *Friends*. Based on the evaluated instances, articles and verb forms are the two most frequent grammar error categories (10 and 8, respectively, out of 23) in DREAM. Besides, the language tends to be less precise in DREAM, indicated by the number of vocabulary suggestions. For example, experts tend to use expressions such as *"really hot"*, *"really beautiful"*,

[2] `https://app.grammarly.com`.

"very bad", and *"very important"* instead of more appropriate yet more advanced adjectives that might hinder reading comprehension of language learners with smaller vocabularies. According to the explanations provided by the tool, the readability scores for both datasets fall into the same category *"Your text is very simple and easy to read, likely to be understood by an average 5th-grader (age 10)"*.

Metric	DREAM	Friends
# of spelling errors	11	146
# of grammar errors	23	16
# of conciseness suggestions	6	2
# of vocabulary suggestions	18	3
General Performance	98.0	95.0
Readability Score	93.7	95.3

Table 4.6: Comparison of the quality of dialogues from DREAM and Friends (a TV show).

4.2 Approaches

We formally introduce the dialogue-based reading comprehension task and notations in Section 4.2.1. To investigate the effects of different kinds of general world knowledge and dialogue structure, we incorporate them into rule-based approaches (Section 4.2.2) as well as non-neural (Section 4.2.3) and neural (Section 4.2.4) machine learning approaches. We describe in detail preprocessing and training in Section 4.2.5.

4.2.1 Problem Formulation and Notations

We start with a formal definition of the dialogue-based multiple-choice reading comprehension task. An n-turn dialogue D is defined as D =

$\{s_1 : t_1, s_2 : t_2, \ldots, s_n : t_n\}$, where s_i represents the speaker ID (e.g., "M" and "W"), and t_i represents the text of the i^{th} turn. Let Q denote the text of question, and $O_{1..3}$ denote the text of three answer options. The task is to choose the correct one from answer options $O_{1..3}$ associated with question Q given dialogue D. In this chapter, we regard this task as a three-class classification problem, each class corresponding to an answer option.

For convenience, we define the following notations, which will be referred in the rest of this chapter. Let D^s denote the turns spoken by speaker s in D. Formally, $D^s = \{s_{i_1} : t_{i_1}, s_{i_2} : t_{i_2}, \ldots, s_{i_m} : t_{i_m}\}$ where $\{i_1, i_2, \ldots, i_m\} = \{i \,|\, s_i = s\}$ and $i_1 < i_2 < \ldots < i_m$. In particular, $s = *$ denotes all the speakers. W^{D^s} and W^{O_i} denote the ordered set of the running words (excluding punctuation marks) in D^s and O_i respectively. Questions designed for dialogue-based reading comprehension often focus on a particular speaker. If there is exactly one speaker mentioned in a question, we use s_Q to denote this target speaker. Otherwise, $s_Q = *$. For example, given the dialogue in Table 4.2, $s_Q =$"M" for Question 1 and 2, and $s_Q = *$ for Question 3.

4.2.2 Rule-Based Approaches

We first attempt to incorporate dialogue structure information into *sliding window* (SW), a rule-based approach developed by Richardson et al. (2013). This approach matches a bag of words constructed from a question Q and one of its answer option O_i with a given document, and calculates the TF-IDF style matching score for each answer option.

Let $\hat{D}^s$, $\hat{Q}$, and $\hat{O}_i$ be the unordered set of distinct words (excluding punctua-

tion marks) in D^s, Q, and O_i, respectively. Instead of only regarding dialogue D as a non-conversational text snippet, we also pay special attention to the context that is relevant to the target speaker mentioned in the question. Therefore, given a target speaker s_Q, we propose to compute a *speaker-focused* sliding window score for each answer option O_i, by matching a bag of words constructed from Q and O_i with D^{s_Q} (i.e., turns spoken by s_Q). Given speaker s, we formally define the sliding window score sw of O_i as:

$$sw_i^s = \max_j \sum_{k=1...|T_i|} \begin{cases} \text{ic}^s(W_{j+k}^{D^s}) & \text{if } W_{j+k}^{D^s} \in T_i \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

where $\text{ic}^s(w) = \log\left(1 + \frac{1}{\sum_i \mathbb{1}(W_i^{D^s}=w)}\right)$, $T_i = \hat{O}_i \cup \hat{Q}$, and $W_i^{D^s}$ denotes the i-th word in W^{D^s}. Based on the above definitions, we can regard sw_i^* as the general score defined in the original sliding window approach, and $sw_i^{s_Q}$ represents the speaker-focused sliding window score considering the target speaker s_Q.

Since sliding window score ignores long-range dependencies, Richardson et al. (2013) introduce a distance-based variation (DSW), in which a word-distance based score is subtracted from the sliding window score to arrive at the final score. Similarly, we calculate the speaker-focused distance-based score given a (Q, O_i) pair and s_Q, by counting the distance between the occurrence of a word in Q and a word in O_i in D^{s_Q}. More formally, given speaker s and a set of stop words[3] U, the distance-based score d of O_i is defined as

$$d_i^s = \begin{cases} 1 & \text{if } |I_Q^s| = 0 \text{ or } |I_{O_i}^s| = 0 \\ \frac{\delta_i^s}{|W^{D^s}|-1} & \text{otherwise} \end{cases} \tag{4.2}$$

[3]We use the list of stop words from NLTK (Bird and Loper, 2004).

where $I^s_Q = (\hat{Q} \cap \hat{D}^s) - U$, $I^s_{O_i} = (\hat{O}_i \cap \hat{D}^s) - \hat{Q} - U$, and δ^s_i is the minimum number of words between an occurrence of a question word and an answer option word in W^{D^s}, plus one. The formal definition of δ^s_i is as follows.

$$\delta^s_i = \min_{W^{D^s}_j \in I^s_Q, W^{D^s}_k \in I^s_{O_i}} |j - k| + 1 \tag{4.3}$$

Based on the above definitions, we can regard d^*_i as the distance-based score defined in the original sliding window approach, and $d^{s_Q}_i$ represents the speaker-focused distance-based score considering speaker s_Q. In addition, the final distance-based sliding window score of O_i (Richardson et al., 2013) can be formulated as

$$sw^*_i - d^*_i \tag{4.4}$$

Compared to (4.4) that only focuses on the general (or speaker-independent) information (i.e., sw^*_i and d^*_i), we can capture general and speaker-focused information (i.e. $sw^{s_Q}_i$ and $d^{s_Q}_i$) simultaneously by averaging them:

$$\frac{sw^{s_Q}_i + sw^*_i}{2} - \frac{d^{s_Q}_i + d^*_i}{2} \tag{4.5}$$

Since a large percentage of questions cannot be solved by word-level matching, we also attempt to incorporate general world knowledge into our rule-based method. We calculate cs^s_i, the maximum cosine similarity between O_i and consecutive words of the same length in W^{D^s}, as:

$$cs^s_i = \max_j \cos\left(\overline{W^{O_i}}, \overline{W^{D^s}_{j \ldots j+|W^{O_i}|-1}}\right) \tag{4.6}$$

where $\overline{x}$ is obtained by averaging the embeddings of the constituent words in x. Here we use ConceptNet embeddings (Speer et al., 2017) since they leverage

the knowledge graph that focuses on general world knowledge. Following (4.5), we capture both general and speaker-focused semantic information within a dialogue as follows.

$$\frac{cs_i^{s_Q} + cs_i^*}{2} \tag{4.7}$$

To make the final answer option selection, our rule-based method combines (4.5) and (4.7):

$$\arg\max_i \frac{sw_i^{s_Q} + sw_i^*}{2} - \frac{d_i^{s_Q} + d_i^*}{2} + \frac{cs_i^{s_Q} + cs_i^*}{2} \tag{4.8}$$

4.2.3 Feature-Based Classifier

To explore what features are effective for dialogue understanding, we first consider a gradient boosting decision tree (GBDT) classifier. Besides the conventional bag-of-words features, we primarily focus on features related to general world knowledge and dialogue structure.

- **Bag of words of each answer option**.
- **Features inspired by rule-based approaches:** we adopt the features introduced in Section 4.2.2, including speaker-independent scores (i.e., sw_i^* and d_i^*) and speaker-focused scores (i.e., $sw_i^{s_Q}$ and $d_i^{s_Q}$).
- **Matching position:** $p_{1..3}^{s_Q}$ and $p_{1..3}^*$, where p_i^s is the last position (in percentage) of a word in D^s that is also mentioned in O_i; 0 if none of the words in D^s is mentioned in O_i. We consider matching position due to our observation of the existence of concessions and negotiations in dialogues (Amgoud et al., 2007). We assume the facts or opinions expressed near the end of a dialogue tend to be more critical for us to answer a question.

- **Pointwise mutual information (PMI):** $pmi^{sQ}_{\max,1..3}$, $pmi^{*}_{\max,1..3}$, $pmi^{sQ}_{\min,1..3}$, $pmi^{*}_{\min,1..3}$, $pmi^{sQ}_{\text{avg},1..3}$, and $pmi^{*}_{\text{avg},1..3}$, where $pmi^{s}_{f,i}$ is defined as

$$pmi^{s}_{f,i} = \frac{\sum_j \log f_k \frac{C_2(W_j^{O_i}, W_k^{D^s})}{C_1(W_j^{O_i})C_1(W_k^{D^s})}}{|W^{O_i}|} \tag{4.9}$$

  $C_1(w)$ denotes the word frequency of w in external copora (we use Reddit posts (Tan and Lee, 2015)), and $C_2(w_1, w_2)$ represents the co-occurrence frequency of word w_1 and w_2 within a distance $< K$ in external copora. We use PMI to evaluate the relatedness between the content of an answer option and the target-speaker-focused context based on co-occurrences of words in external corpora, inspired by previous studies on narrative event chains (Chambers and Jurafsky, 2008).

- **ConceptNet relations (CR):** $cr_{1..3,1..|R|}$. $R = \{r_1, r_2, \ldots\}$ is the set of ConceptNet relation types (e.g., *"CapableOf"* and *"PartOf"*). $cr_{i,j}$ is the number of relation triples (w_1, r_j, w_2) that appear in the ConceptNet (Speer et al., 2017), where w_1 represents a word in answer option O_i, w_2 represents a word in D, and the relation type $r_j \in R$. Similar to the motivation of using PMI, we use CR to capture the association between an answer option and the source dialogue based on raw co-occurrence counts in the commonsense knowledge base.

- **ConceptNet embeddings (CE):** besides the lexical similarity based on string matching, we also calculate $cs^{*}_{1..3}$ and $cs^{sQ}_{1..3}$, where cs^{*}_{i} and cs^{sQ}_{i} represent the maximum cosine similarity between O_i and consecutive words of the same length in D and D^{sQ}, respectively (Expression 4.6 in Section 4.2.2). We use ConceptNet embeddings (Speer et al., 2017) since they leverage the general world knowledge graph.

4.2.4 End-To-End Neural Network

Our end-to-end neural model is based on a generative pre-trained language model (LM). We follow the framework of finetuned transformer LM (FTLM) (Radford et al., 2018) and make modifications for dialogue-based reading comprehension.

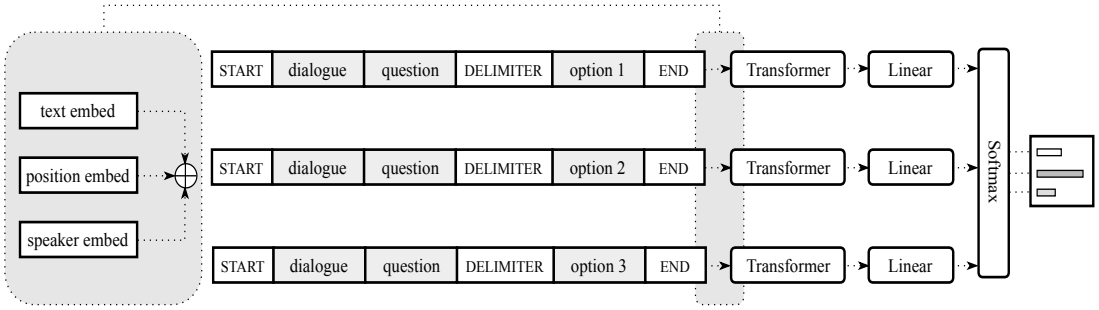


Figure 4.1: Overall neural network framework (Section 4.2.4).

The training procedure of FTLM consists of two stages. The first stage is to learn a high-capacity language model on a large-scale unsupervised corpus of tokens $\mathcal{U} = \{u_1, \ldots, u_n\}$ by maximizing the following likelihood:

$$L_{LM}(\mathcal{U}) = \sum_i \log P(u_i \mid u_{i-k}, \ldots, u_{i-1}; \Theta) \tag{4.10}$$

where k is the context window size, and the conditional probability P is modeled by a multi-layer transformer decoder (Liu et al., 2018) with parameters Θ. In the second stage, the model is adapted to a labeled dataset C, where each instance consists of a sequence of input tokens $x^1, \ldots, x^m$ with a label y, by maximizing:

$$L(C) = \sum_{x,y} \log P(y \mid x^1, \ldots, x^m) + \lambda L_{LM}(C) \tag{4.11}$$

where $P(y \mid x^1, \ldots, x^m)$ is obtained by a linear + softmax layer over the final transformer block's activation, and λ is the weight for language model. For multiple-choice reading comprehension, the input tokens $x^1, \ldots, x^m$ come from the con-

catenation of a start token, dialogue, question, a delimiter token, answer option, and an end token; y indicates if the answer option is correct.

Since the original FTLM framework already leverages rich linguistic information from a large unlabeled corpus, which can be regarded as a type of tacit general world knowledge, we investigate whether additional dialogue structure can further improve this strong baseline. We propose *speaker embedding* to better capture dialogue structure. Specifically, in the original framework, given an input context $(u_{-k}, \ldots, u_{-1})$ of the transformer, the encoding of u_{-i} is $\boldsymbol{we}(u_{-i}) + \boldsymbol{pe}(i)$, where $\boldsymbol{we}(\cdot)$ is the word embedding, and $\boldsymbol{pe}(\cdot)$ is the position embedding. When adapting Θ to DREAM, we change the encoding to $\boldsymbol{we}(u_{-i}) + \boldsymbol{pe}(i) + \boldsymbol{se}(u_{-i}, s_Q)$ where the speaker embedding $\boldsymbol{se}(u_{-i}, s_Q)$ is (a) $\mathbf{0}$ if the token u_{-i} is not in the dialogue (i.e. it is either a start/end/delimiter token or a token in the question/option); (b) $\boldsymbol{e}_{target}$ if the token is spoken by s_Q; (c) $\boldsymbol{e}_{rest}$ if the token is in the dialogue but not spoken by s_Q. $\boldsymbol{e}_{target}$ and $\boldsymbol{e}_{rest}$ are trainable and initialized randomly. We show the overall framework in Figure 4.1.

4.2.5 Preprocessing and Training Details

For all the models, we conduct coreference resolution to determine speaker mentions of s_Q based on simple heuristics. Particularly, we map three most common speaker abbreviations (i.e., *"M"*; *"W"* and *"F"*) that appear in dialogues to their eight most common corresponding mentions (i.e., *"man"*, *"boy"*, *"he"*, and *"his"*; *"woman"*, *"girl"*, *"she"*, and *"her"*) in questions. We keep speaker abbreviations unchanged, since neither replacing them with their corresponding full forms nor removing them contributes to the performance based on our

experiments.

For the neural model mentioned in Section 4.2.4, most of our parameter settings follow Radford et al. (2018). We adopt the same preprocessing procedure and use their publicly released language model, which is pre-trained on the BooksCorpus dataset (Zhu et al., 2015). We set the batch size to 8, language model weight λ to 2, and maximum epochs of training to 10.

For other models, we use the following preprocessing steps. We tokenize and lowercase the corpus, convert number words to numeric digits, normalize time expressions to 24-hour numeric form, and address negation by removing interrogative sentences that receive *"no"* as the reply. We use the gradient boosting classifier implemented in the scikit-learn toolkit (Pedregosa et al., 2011). We set the number of boosting iterations to 600 and keep the rest of hyperparameters unchanged. The distance upper bound K for PMI is set to 10.

We perform several runs of machine learning models (Section 4.2.3 and Section 4.2.4) with randomness introduced by different random seeds and/or GPU non-determinism and select the model or models (for ensemble) that perform best on the development set.

4.3 Experiment

4.3.1 Baselines

We implement several baselines, including rule-based methods and state-of-the-art neural models.

Method	Dev	Test
Random	32.8	33.4
Word Matching (WM) (Yih et al., 2013)	41.7	42.0
Sliding Window (SW) (Richardson et al., 2013)	42.6	42.5
Distance-Based Sliding Window (DSW) (Richardson et al., 2013)	44.4	44.6
Stanford Attentive Reader (SAR) (Chen et al., 2016)	40.2	39.8
Gated-Attention Reader (GAR) (Dhingra et al., 2017)	40.5	41.3
Co-Matching (CO) (Wang et al., 2018d)	45.6	45.5
Finetuned Transformer LM (FTLM) (Radford et al., 2018)	55.9	55.5
Our Approaches:		
DSW++ (DSW w/ Dialogue Structure and ConceptNet Embedding)	51.4	50.1
GBDT++ (GBDT w/ Features of Dialogue Structure and General World Knowledge)	53.3	52.8
FTLM++ (FTLM w/ Speaker Embedding)	**57.6**	**57.4**
Ensemble of 3 FTLM++	58.1	58.2
Ensemble of 1 GBDT++ and 3 FTLM++	**59.6**	**59.5**
Human Performance	93.9⋆	95.5⋆
Ceiling Performance	98.7⋆	98.6⋆

Table 4.7: Performance in accuracy (%) on the DREAM dataset. Performance marked by ⋆ is reported based on 25% annotated questions from the development and test sets.

- **Word Matching** This strong baseline (Yih et al., 2013) selects the answer option that has the highest count of overlapping words with the given dialogue.
- **Sliding Window** We implement the sliding window approach (i.e., $\arg\max_i sw_i^*$) and its distance-based variation DSW (i.e., $\arg\max_i sw_i^* - d_i^*$) (Richardson et al., 2013) introduced in Section 4.2.2.
- **Enhanced Distance-Based Sliding Window (DSW++)** We also use general world knowledge and speaker-focused information to improve the original sliding window baseline, formulated in Expression 4.8 (Section 4.2.2).
- **Stanford Attentive Reader** This neural baseline compares each candidate answer (i.e., entity) representation to the question-aware document representation built with attention mechanism (Hermann et al., 2015; Chen et al., 2016). Lai et al. (2017) add a bilinear operation to compare document and answer option representations to answer multiple-choice questions.

- **Gated-Attention Reader** The baseline models multiplicative question-specific document representations based on a gated-attention mechanism (Dhingra et al., 2017), which are then compared to each answer option (Lai et al., 2017).
- **Co-Matching** This state-of-the-art multiple-choice reading comprehension model explicitly treats question and answer option as two sequences and jointly matches them against a given document (Wang et al., 2018d).
- **Finetuned Transformer LM** This is a general task-agnostic model introduced in Section 4.2.4, which achieves the best reported performance on several tasks requiring multi-sentence reasoning (Radford et al., 2018).

We do not investigate other ways of leveraging pre-trained deep models such as adding ELMo representations (Peters et al., 2018) as additional features to a neural model since recent studies show that directly fine-tuning a pre-trained language model such as FTLM is significantly superior on multiple-choice reading comprehension tasks (Radford et al., 2018; Chen et al., 2019). We do not apply more recent extractive models such as AOA (Cui et al., 2017) and QANet (Yu et al., 2018) since they aim at precisely locating a span in a document. When adapted to solve questions with abstractive answer options, extractive models generally tend to perform less well (Chen et al., 2016; Dhingra et al., 2017; Lai et al., 2017).

4.3.2 Results and Analysis

We report the performance of the baselines introduced in Section 4.3.1 and our proposed approaches in Table 4.7. We report the averaged accuracy of two

annotators as the human performance. The proportion of valid questions (i.e., an unambiguous question with a unique correct answer option provided) that are manually checked by annotators on the annotated test and development sets is regarded as the human ceiling performance.

Surface matching is insufficient. Experimental results show that neural models that primarily exploit surface-level information (i.e., SAR, GAR, and CO) attain a performance level close to that of simple rule-based approaches (i.e., WM, SW, and DSW). The highest accuracy achieved by CO is 45.5%, a similar level of performance to the rule-based method DSW (44.6%).

It is helpful to incorporate general world knowledge and dialogue structure. We see a significant gain 5.5% in accuracy when enhancing DSW using general world knowledge from ConceptNet embeddings and considering speaker-focused information (Section 4.2.2). FTLM, which leverages rich external linguistic knowledge from thousands of books, already achieves a much higher accuracy 55.5% compared to previous state-of-the-art machine comprehension models, indicating the effectiveness of general world knowledge. Experimental results show that our best single model FTLM++ significantly outperforms FTLM (p-value = 0.03), illustrating the usefulness of additional dialogue structure. Compared to the state-of-the-art neural reader Co-Matching that primarily explores surface-level information (45.5%), the tacit general world knowledge (in the pre-trained language model) and dialogues structure in FTLM++ lead to an absolute gain of 11.9% in accuracy.

Ensembling different types of methods can bring further improvements. We employ the majority vote strategy to obtain the ensemble model performance. While GBDT++ (52.8%) itself does not outperform FTLM++, GBDT++ can serve

as a supplement to FTLM++ as they leverage different types of general world knowledge and model architectures. We achieve the highest accuracy 59.5% by ensembling one GBDT++ and three FTLM++.

4.3.3 Ablation Tests

We conduct ablation tests to evaluate the individual components of our proposed approaches (Table 4.8). In Table 4.9, we summarize the involved types of dialogue structure and general world knowledge in our approaches.

Dialogue Structure Specifically, we observe 1.4% drop in accuracy if we set the target speaker s_Q to $*$ for all questions when we apply DSW++. We observe a similar performance drop when we remove speaker-focused features from GBDT++. In addition, removing speaker embeddings from FTLM++ leads to 1.7% drop in accuracy (in this case, the model becomes the original FTLM). These results consistently indicate the usefulness of dialogue structure for dialogue understanding.

General World Knowledge We also investigate the effects of general world knowledge. The accuracy of DSW++ drops by 4.7% if we remove ConceptNet embeddings (CE) by deleting the last term of Expression 4.8 in Section 4.2.2. Additionally, the accuracy of GBDT++ drops by 6.2% if we remove all the general world knowledge features (i.e., ConceptNet embeddings/relations and PMI), leading to prediction failures on questions such as *"What do we learn about the man?"* whose correct answer option *"He is health-conscious."* is not explicitly mentioned in the source dialogue *"**M**: We had better start to eat onions frequently, Linda. **W**: But you hate onions, don't you? **M**: Until I learned from a report from today's*

paper that they protect people from flu and colds. After all, compared with health, taste is not so important.". Moreover, if we train FTLM++ with randomly initialized transformer weights instead of weights pre-trained on the external corpus, the accuracy drops dramatically to 36.2%, which is only slightly better than a random baseline.

Method	**Accuracy**	**Δ**
DSW++	51.4	–
– dialogue structure	50.0	-1.4
– CE	46.7	-4.7
GBDT++	53.3	–
– bag of words	51.6	-1.7
– rule-based features	51.2	-2.1
– matching position	53.0	-0.3
– dialogue structure	51.9	-1.4
– PMI	51.4	-1.9
– CR	52.7	-0.6
– CE	52.7	-0.6
– PMI, CR, CE	47.1	-6.2
FTLM++	57.6	–
– speaker embedding	55.9	-1.7
– LM pre-training	36.2	-21.4

Table 4.8: Ablation tests on the development set (%). Minus (−) indicates percentage decrease.

	Dialogue Structure	General World Knowledge
DSW++	speaker-focused scores	CE
GBDT++	speaker-focused features	CE, CR, and PMI
FTLM++	speaker embedding	pre-trained LM

Table 4.9: Types of dialogue structure and general world knowledge investigated in our approaches.

4.3.4 Error Analysis

Impact of Longer Turns The number of dialogue turns has a significant impact on the performance of FTLM++. As shown in Figure 4.2, its performance reaches

the peak while the number of turns ranges from 0 to 10 while it suffers severe performance drops when the given dialogue contains more turns. Both DSW++ (56.8%) and GBDT++ (57.4%) outperform FTLM++ (55.7%) when the number of turns ranges from 10 to 48. To deal with lengthy context, it may be helpful to first identify relevant sentences based on a question and its associated answer options rather than using the entire dialogue context as input.

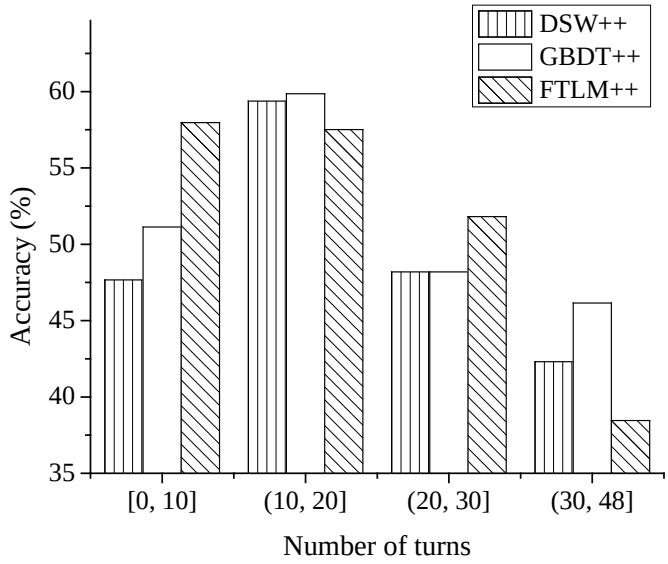


Figure 4.2: Performance comparison of different number of turns on the test set.

Impact of Confusing Distractors For 54.5% of questions on the development set, the fuzzy matching score (Sikes, 2007) of at least one distractor answer option against the dialogue is higher than the score of the correct answer option. For questions that all models (i.e., DSW++, GBDT++, and FTLM++) fail to answer correctly, 73.0% of them contain at least one such confusing distractor answer option. The causes of this kind of errors can be roughly divided into two categories. First, the distractor is wrongly associated with the target speaker/s mentioned in the question (e.g., answer option A and C in D2-Q3 in Table 4.2). Second, although the claim in the distractor is supported by the dialogue, it is irrelevant to

the question (e.g., D1-Q1-B in Table 4.1). A promising direction to solve this problem could be the construction of speaker-focused event chains (Chambers and Jurafsky, 2008) and advanced dialogue-specific coreference resolution systems for more reliable evidence context detection in a dialogue.

Impact of Question Types We further report the performance of the best single model FTLM++ and the GBDT++ baseline on the categories defined in Section 4.1.2 (Table 4.10). Not surprisingly, both models perform worse than random guessing on math problems. While most of the math problems can be solved by one single linear equation, it is still difficult to apply recent neural math word problem solvers (Huang et al., 2018; Wang et al., 2018a) due to informal dialogue-based problem descriptions and the requirement of commonsense inference. For example, given the dialogue:

"***W**: The plane arrives at 10:50. It is already 10:40 now. Be quick!* ***M**: Relax. Your watch must be fast. There are still twenty minutes left.*",

we need prior knowledge to infer that the watch of the man is showing incorrect time 10:40. Instead, 10:50 should be used as the reference time with the time interval *"twenty minutes left"* together to answer the question *"What time is it now?"*.

Results show that GBDT++ is superior to the fine-tuned language model on the questions under the category *matching* (68.1% vs. 57.0%) and the latter model is more capable of answering implicit questions (e.g., under the category *summary*, *logic*, and *commonsense*) which require aggregation of information from multiple sentences, the understanding of the entire dialogue, or the utilization of world knowledge. Therefore, it might be useful to leverage the strengths of individual models to solve different types of questions.

Question Type	FTLM++	GBDT++
Matching	57.0	**68.1**
Reasoning	**56.8**	49.4
Summary	**73.6**	47.1
Logic	**55.0**	49.7
Arithmetic	**30.2**	24.5
Commonsense	**53.4**	41.7
Single sentence	56.5	**63.3**
Multiple sentences	**56.9**	49.5

Table 4.10: Accuracy (%) by question type on the annotated development subset.

4.4 Related Work

We divide reading comprehension datasets into three categories based on the types of answers.

	SQuAD	NarrativeQA	CoQA	RACE	DREAM (this work)
Answer type	extractive	abstractive	abstractive	multiple-choice	multiple-choice
Source document type	written text	written text	written text	written text	dialogue
# of source documents	536	1,572	8,399	27,933	6,444
Average answer length	3.2	4.7	2.7	5.3	5.3
Extractive (%)	100.0	73.6	66.8	13.0	16.3
Abstractive (%)	0.0	26.4	33.2	87.0	83.7

Table 4.11: Distribution of answer (or correct answer option) types in three kinds of reading comprehension datasets. Statistics of other datasets come from Reddy et al. (2019), Kočiskỳ et al. (2018), and Lai et al. (2017).

4.4.1 Extractive and Abstractive Datasets

In recent years, we have seen increased interest in large-scale cloze/span-based reading comprehension dataset construction (Hermann et al., 2015; Hill et al., 2016; Onishi et al., 2016; Rajpurkar et al., 2016; Bajgar et al., 2016; Nguyen et al., 2016; Trischler et al., 2017; Joshi et al., 2017; Choi et al., 2018). We regard them as extractive since candidate answers are usually short spans from source docu-

ments. State-of-the-art neural models with attention mechanisms already achieve very high performance based on local lexical information. Recently researchers work on the construction of spoken span-based datasets (Lee et al., 2018; Li et al., 2018a) by applying text-to-speech technologies or recruiting human speakers based on formal written document-based datasets such as **SQuAD** (Rajpurkar et al., 2016). Some span-based conversation datasets are constructed from a relatively small size of dialogues from TV shows (Chen and Choi, 2016; Ma et al., 2018).

Considering the limitations in extractive datasets, answers in abstractive datasets such as MS MARCO (Nguyen et al., 2016), SearchQA (Dunn et al., 2017), and **NarrativeQA** (Kočiskỳ et al., 2018) are human crowdsourced based on source documents or summaries. Concurrently, there is a growing interest in conversational reading comprehension such as **CoQA** (Reddy et al., 2019). Since annotators tend to copy spans as answers (Reddy et al., 2019), the majority of answers are still extractive in these datasets (Table 4.11). Compared to the datasets mentioned above, most of the correct answer options (**83.7%**) in DREAM are free-form text.

4.4.2 Multiple-Choice Datasets

We primarily discuss the multiple-choice datasets in which answer options are not restricted to extractive text spans in the given document. Instead, most of the correct answer options are abstractive (Table 4.11). Multiple-choice datasets involve extensive human involvement for problem generation during crowdsourcing (i.e., questions, correct answer option, and distractors). Besides surface

matching, a significant portion of questions require multiple-sentence reasoning and external knowledge (Richardson et al., 2013; Mostafazadeh et al., 2016; Khashabi et al., 2018; Ostermann et al., 2018).

Besides crowdsourcing, some datasets are collected from examinations designed by educational experts (Penas et al., 2014; Shibuki et al., 2014; Tseng et al., 2016; Clark et al., 2016; Lai et al., 2017; Mihaylov et al., 2018), which aim to test human examinees. There are various types of complicated questions such as math word problems, summarization, logical reasoning, and sentiment analysis. Since we can adopt more objective evaluation criteria such as accuracy, these questions are usually easy to grade. Besides, questions from examinations are generally clean and high-quality. Therefore, human performance ceiling on this kind of datasets is much higher (e.g., 94.5% on **RACE** (Lai et al., 2017) and 98.6% on DREAM in accuracy) than that of datasets built by crowdsourcing.

In comparison, we present the first multiple-choice **dialogue-based** dataset from examinations that contains a large percentage of questions that require multiple sentence inference. To the best of our knowledge, DREAM also contains the largest number of questions involving **commonsense reasoning** compared to other examination datasets.

4.5 Chapter Summary

We present DREAM, the first multiple-choice dialogue-based reading comprehension dataset from English language examinations. Besides the multi-turn multi-party dialogue context, 85% of questions require multiple-sentence reasoning, and 34% of questions also require commonsense knowledge, making this

task very challenging. We apply several popular reading comprehension models and find that surface-level information is insufficient. We incorporate general world knowledge and dialogue structure into rule-based and machine learning methods and show the effectiveness of these factors, suggesting a promising direction for dialogue-based reading comprehension.

CHAPTER 5

INVESTIGATING PRIOR KNOWLEDGE FOR CHINESE READING COMPREHENSION

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

Adams and Bruce (1982)

In the previous two chapters, we present our efforts in the development of techniques tackling a variety of free-form multiple-choice MRC tasks that contain a significant percentage of questions focusing on the implicitly expressed facts, events, opinions, or emotions in the given text (Richardson et al., 2013; Lai et al., 2017; Ostermann et al., 2018; Khashabi et al., 2018; Sun et al., 2019a). Generally, we require the integration of our own prior knowledge and the information presented in the given text to answer these questions, posing significant challenges for MRC systems. However, until recently, progress in the development of techniques for addressing this kind of MRC task for Chinese has lagged behind their English counterparts. A primary reason is that most previous work focuses on constructing MRC datasets for Chinese in which most answers are either spans (Cui et al., 2016; Li et al., 2016; Cui et al., 2018a; Shao et al., 2018) or abstractive texts (He et al., 2017) merely based on the information **explicitly** expressed in the provided text.

With a goal of developing similarly challenging, but free-form multiple-choice datasets, and promoting the development of MRC techniques for Chinese, we

introduce in this chapter the first free-form multiple-**C**hoice **C**hinese machine reading **C**omprehension dataset (C^3) 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, C^3 includes a dialogue-based task C^3_D in which the given document is a **d**ialogue, and a mixed-genre task C^3_M in which the given document is a **m**ixed-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 C^3 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 C^3) 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 to tasks accompanied with relatively formally written texts. Specifically, compared to existing MRC datasets for Chinese (He et al., 2017; Cui et al., 2018b), C^3 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 C^3 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.

This chapter is based on Sun et al. (2020b).

5.1 Data

In this section, we describe the construction of C^3 (Section 5.1.1). We also analyze the data (Section 5.1.2) and the types of prior knowledge needed for the MRC tasks (Section 5.1.3).

5.1.1 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

1928年，经徐志摩介绍，时任中国公学校长的胡适聘用了沈从文做讲师，主讲大学一年级的现代文学选修课。 当时，沈从文已经在文坛上崭露头角，在社会上也小有名气，因此还未到上课时间，教室里就坐满了学生。上课时间到了，沈从文走进教室，看见下面黑压压一片，心里陡然一惊，脑子里变得一片空白，连准备了无数遍的第一句话都堵在嗓子里说不出来了。 他呆呆地站在那里，面色尴尬至极，双手拧来拧去无处可放。上课前他自以为成竹在胸，所以就没带教案和教材。整整10分钟，教室里鸦雀无声，所有的学生都好奇地等着这位新来的老师开口。沈从文深吸了一口气，慢慢平静了下来，原先准备好的东西也重新在脑子里聚拢，然后他开始讲课了。不过由于他依然很紧张，原本预计一小时的授课内容，竟然用了不到15分钟就讲完了。 接下来怎么办？他再次陷入了窘境。无奈之下，他只好拿起粉笔在黑板上写道：我第一次上课，见你们人多，怕了。 顿时，教室里爆发出了一阵善意的笑声，随即一阵鼓励的掌声响起。得知这件事之后，胡适对沈从文大加赞赏，认为他非常成功。有了这次经历，在以后的课堂上，沈从文都会告诫自己不要紧张，渐渐地，他开始在课堂上变得从容起来。	In 1928, recommended by Hsu Chih-Mo, Hu Shih, who was the president of the previous National University of China, employed Shen Ts'ung-wen as a lecturer of the university in charge of teaching the optional course of modern literature. At that time, Shen already made himself conspicuous in the literary world and was a little famous in society. For this sake, 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 believed 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. Nevertheless, 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 again 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. Hearing this episode, Hu heaped praise upon Shen, thinking that he was very successful. Because of this experience, Shen always reminded himself of not being nervous in his class for years afterwards. Gradually, he began to give his lecture at leisure in class.
Q1 第2段中，“黑压压一片”指的是： A. 教室很暗 B. 听课的人多★ C. 房间里很吵 D. 学生们发言很积极	**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 沈从文没拿教材，是因为他觉得： A. 讲课内容不多 B. 自己准备得很充分★ C. 这样可以减轻压力 D. 教材会限制自己的发挥	**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 看见沈从文写的那句话，学生们： A. 急忙安慰他 B. 在心里埋怨他 C. 受到了极大的鼓舞 D. 表示理解并鼓励了他★	**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 上文主要谈的是： A. 中国教育制度的发展 B. 紧张时应如何调整自己 C. 沈从文第一次讲课时的情景★ D. 沈从文如何从作家转变为教师的	**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 5.1: A C^3_M problem and its English translation (★: the correct option).

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.
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 wait in line 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.

Table 5.2: English translation of a sample problem from C^3_D (⋆: the correct option).

Metric	$\mathbf{C^3_M}$	$\mathbf{C^3_D}$	$\mathbf{C^3}$
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. option length (in characters)	6.5 / 45	4.4 / 31	5.5 / 45
Avg./Max. question 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
non-extractive correct option (%)	81.9	78.9	80.4
# of documents / # of questions			
Training	3,138 / 6,013	4,885 / 5,856	8,023 / 11,869
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 5.3: The overall statistics of C^3. $C^3 = C^3_M \cup C^3_D$.

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 5.1 and 5.2, 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.

5.1.2 Data Statistics

We summarize the overall statistics of C^3 in Table 5.3. We observe notable differences exist between C^3_M and C^3_D. For example, C^3_M, in which most documents are formally written texts, has a larger vocabulary size compared to that of C^3_D 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 C^3_M is longer than that in C^3_D (76.3). In general, C^3 may not be suitable for evaluating the comprehension ability of machine readers on lengthy texts

as the average length of document C^3 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).

	C^3_M	C^3_D	C^3	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	2.0	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	5.0	2.7	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

Table 5.4: 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.

5.1.3 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 C^3 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 5.1 and Q3 in Table 5.2), 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; 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 5.1 and Q2 in Table 5.2 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 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 5.1 in this category as *"friendly laughter"* may express *"understanding"*. Q1 in Table 5.2 about the relation between the two speakers also belongs to

this category.

- Precondition[†]: If had event A 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 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 C^3 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 dataset RACE (Lai et al., 2017) and DREAM (Chapter 4) that can be regarded as the English counterpart of C^3_M and C^3_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.

C^3_M vs. C^3_D As shown in Table 5.4, compared to the dialogue-based task (C^3_D), C^3_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 C^3_D need scenario knowledge perhaps due to 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.

C^3 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, C^3_D and its English counterpart DREAM (Chapter 4) have similar problem formats, document types, and data collection methodologies (from Chinese-as-a-second-language and English-as-a-foreign-language exams, respectively). We notice that the knowledge type distributions of the two datasets are indeed very similar. Therefore, C^3 may facilitate future cross-lingual MRC studies.

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; Kočiskỳ et al., 2018; Reddy et al., 2019). Namely, turkers asked to write answers in his/her 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.

Method	C^3_M		C^3_D		C^3	
	Dev	**Test**	**Dev**	**Test**	**Dev**	**Test**
Random	27.8	27.8	26.4	26.6	27.1	27.2
Distance-Based Sliding Window (Richardson et al., 2013)	47.9	45.8	39.6	40.4	43.8	43.1
Co-Matching (Wang et al., 2018d)	47.0	48.2	55.5	51.4	51.0	49.8
BERT (Devlin et al., 2019)	65.6	64.6	65.9	64.4	65.7	64.5
ERNIE (Sun et al., 2019d)	63.7	63.6	67.3	64.6	65.5	64.1
BERT-wwm (Cui et al., 2019)	66.1	64.0	64.8	65.0	65.5	64.5
BERT-wwm-ext (Cui et al., 2019)	67.9	68.0	67.7	68.9	67.8	68.5
Human Performance*	96.0	93.3	98.0	98.7	97.0	96.0

Table 5.5: Performance of baseline in accuracy (%) on the C^3 dataset ($*$: based on the annotated subset of test and development sets of C^3).

5.2 Approaches

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

5.2.1 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.

5.2.2 Co-Matching

We employ Co-Matching (Wang et al., 2018d), 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 5.3.1).

5.2.3 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., 2019d), 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.

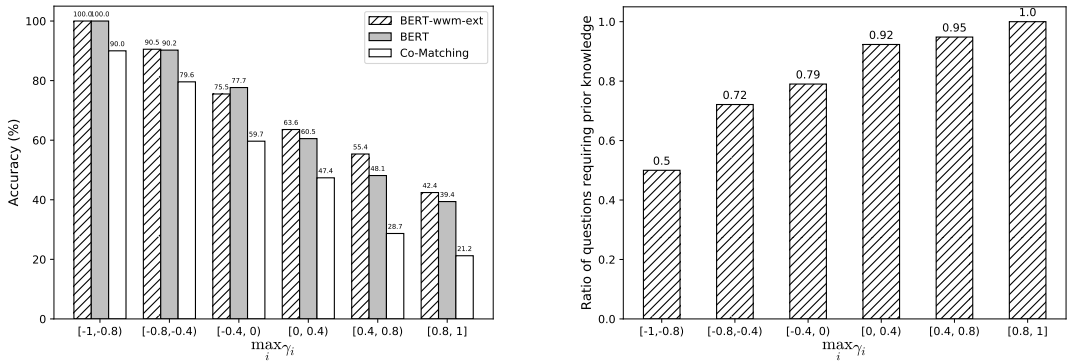
Given document d, question q, and answer option o_i, we construct the input sequence by concatenating `[CLS]`, tokens in d, `[SEP]`, tokens in q, `[SEP]`, tokens in o_i, and `[SEP]`, where `[CLS]` and `[SEP]` are the classifier token and sentence separator in a pre-trained language model, respectively. We add an embedding

vector $\boldsymbol{t_1}$ to each token before the first [SEP] (inclusive) and an embedding vector $\boldsymbol{t_2}$ to every other token, where $\boldsymbol{t_1}$ and $\boldsymbol{t_2}$ 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.

5.3 Experiment

	Co-Matching C^3_M— C^3_D	**BERT** C^3_M— C^3_D	**BERT-wwm-ext** C^3_M— C^3_D	**Human** C^3_M— C^3_D
Matching	54.6 — 70.4	81.8 — 81.5	100.0 — 85.2	100.0 — 100.0
Prior knowledge	47.5 — 51.2	64.0 — 64.2	62.6 — 68.3	95.7 — 97.6
⋄ Linguistic	49.4 — 49.0	67.1 — 62.8	61.2 — 68.6	97.7 — 100.0
⋄ Domain-specific*	– — 66.7	– — 0.0	– — 0.0	– — 100.0
⋄ General world	46.5 — 53.8	57.7 — 66.3	64.8 — 70.0	93.0 — 96.3
Arithmetic*	50.0 — 60.0	0.0 — 80.0	50.0 — 60.0	100.0 — 100.0
Connotation*	0.0 — 50.0	0.0 — 62.5	0.0 — 62.5	100.0 — 100.0
Cause-effect	47.6 — 55.6	57.1 — 55.6	66.7 — 66.7	95.2 — 100.0
Implication	46.7 — 45.5	70.0 — 50.0	70.0 — 54.6	86.7 — 95.5
Part-whole	60.0 — 50.0	40.0 — 50.0	40.0 — 50.0	100.0 — 83.3
Precondition*	66.7 — 50.0	66.7 — 25.0	66.7 — 75.0	100.0 — 100.0
Scenario	40.0 — 61.3	40.0 — 80.7	60.0 — 83.9	100.0 — 96.8
Other*	– — 0.0	– — 0.0	– — 0.0	– — 100.0
Single sentence	50.0 — 64.7	72.4 — 76.5	71.1 — 82.4	97.4 — 97.1
Multiple sentences	47.2 — 51.7	58.3 — 64.7	61.1 — 68.1	94.4 — 98.3
Independent*	0.0 — –	50.0 — –	0.0 — –	100.0 — –

Table 5.6: Performance comparison in accuracy (%) by categories based on a subset of development sets of C^3 ($*$: ≤ 10 annotated instances fall into that category).



(a) Performance comparison based on different largest distractor plausibility.

(b) Correlation between largest distractor plausibility and the need for prior knowledge.

Figure 5.1: Analysis of distractor plausibility.

5.3.1 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 employ 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. (2018b).

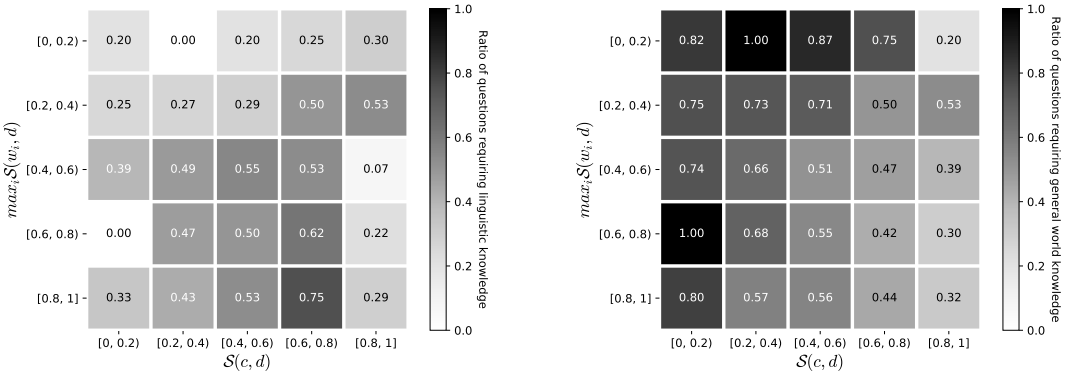
[1] `https://github.com/hankcs/HanLP`.

Fine-Tuning Pre-Trained Language Models. We set the learning rate, batch size, and maximal sequence length to 2×10^{-5}, 24, and 512, respectively. We truncate the longest sequence among d, q, and o_i (Section 5.2.3) 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).

5.3.2 Baseline Results

As shown in Table 5.5, 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 to 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 5.6), 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 5.1.3).



(a) The need for linguistic knowledge.

(b) The need for general world knowledge.

Figure 5.2: The need for two major types of prior knowledge when answering questions of different $\max_i \mathcal{S}(w_i, d)$ and $\mathcal{S}(c, d)$.

5.3.3 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:

$$\gamma_i = \mathcal{S}(w_i, d) - \mathcal{S}(c, d) \tag{5.1}$$

where $\mathcal{S}(x, y)$ 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, $\mathcal{S}(x, y) = 1$; if x shares no character with y, $\mathcal{S}(x, y) = 0$. By definition, $\mathcal{S}(w_i, d)$ in Equation (5.1) measures the lexical similarity between distractor w_i and d; $\mathcal{S}(c, d)$ 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 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 5.1(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 on the annotated instances, as shown in Figure 5.1(b). For further analysis, we group annotated instances by different $\max_i \mathcal{S}(w_i, d)$ and $\mathcal{S}(c, d)$ (in Equation (5.1)) and separately compare their need for linguistic knowledge and general world knowledge. As shown in Figure 5.2, general world knowledge is crucial for question

answering when the correct answer option is not mentioned explicitly in the document (i.e., $\mathcal{S}(c, d)$ is relatively small). In contrast, we tend to require linguistic knowledge when both the correct answer option and the most confusing distractor (i.e., the one with the largest distractor plausibility) are very similar to the given document.

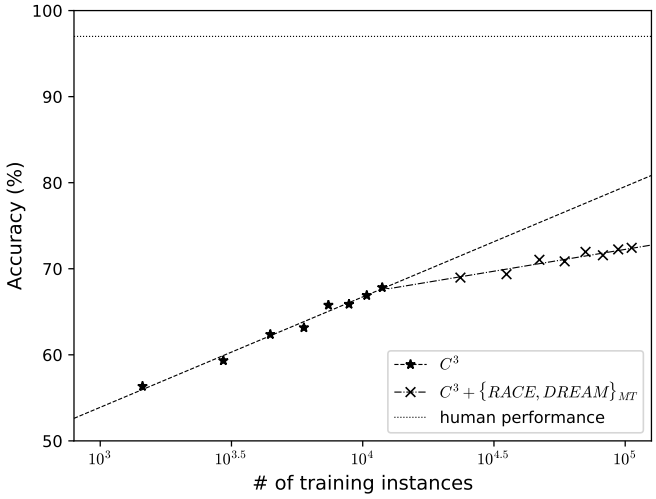


Figure 5.3: Performance of BERT-wwm-ext trained on $1/8, 2/8, \ldots, 8/8$ of C^3 training data, and C^3 training data plus $1/8, 2/8, \ldots, 8/8$ of machine translated (MT) RACE and DREAM training data.

5.3.4 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 C^3. As shown in Figure 5.3, 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 C^3 and its English counterparts RACE and DREAM are highly similar (Section 5.1.3), we translate RACE and DREAM from English to Chinese by Google Translate and plot the performance of BERT-wwm-ext trained on C^3 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 C^3. From another perspective, BERT-wwm-ext trained on all translated instances **without** using any data in C^3 only achieves an accuracy of 67.1% on the development set of C^3, slightly worse than 67.8% achieved when only the training data in C^3 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 C^3 and all *original* training instances in English from RACE and DREAM. However, the accuracy on the development set of C^3 is 63.4%, which is even lower than the performance (65.7% in Table 5.5) of fine-tuning Chinese BERT-Base only on C^3.

5.4 Related Work

We will first discuss standard MRC datasets for English, followed by MRC/QA datasets for Chinese.

Chinese Task	Document Genre	Question Type	Answer Type	Question Size	English Counterpart
Question Answering					
QS (Cheng et al., 2016)	N/A	free-form	multiple-choice	0.6K	ARC (Clark et al., 2016)
MCQA (Guo et al., 2017a)	N/A	free-form	multiple-choice	14.4K	ARC (Clark et al., 2016)
MedQA (Zhang et al., 2018b)	N/A	free-form	multiple-choice	235.2K	ARC (Clark et al., 2016)
GeoSQA (Huang et al., 2019b)	N/A	free-form	multiple-choice	4.1K	DD (Lally et al., 2017)
Machine Reading Comprehension					
PD (Cui et al., 2016)	news	cloze	extractive	876.7K	CNN/Daily (Hermann et al., 2015)
CFT (Cui et al., 2016)	books	cloze	extractive	3.6K	CBT (Hill et al., 2016)
CMRC 2018 (Cui et al., 2018b)	Wiki	free-form	extractive	19.1K	SQuAD (Rajpurkar et al., 2016)
DuReader (He et al., 2017)	web	free-form	abstractive	≈ 200K	MS MARCO (Nguyen et al., 2016)
ChID (Zheng et al., 2019)	mixed-genre	cloze	multiple-choice	728.7K	CLOTH (Xie et al., 2018)
$\mathbf{C^3_M}$ **(this work)**	mixed-genre	free-form	multiple-choice	10.0K	RACE (Lai et al., 2017)
$\mathbf{C^3_D}$ **(this work)**	dialogue	free-form	multiple-choice	9.6K	DREAM (Sun et al., 2019a)

Table 5.7: Comparison of C^3 and representative Chinese question answering and machine reading comprehension tasks. We list only one English counterpart for each Chinese dataset.

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; Kočiskỳ 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 datasets 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 5.1.3.

Chinese. Extractive MRC datasets for Chinese (Cui et al., 2016; Li et al., 2016; Cui et al., 2018b,a; 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., 2019b) 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 5.7.

5.5 Chapter Summary

We present the first free-form multiple-choice Chinese machine reading comprehension dataset (C^3), 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 that 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.

CHAPTER 6

IMPROVING READING COMPREHENSION WITH CONTEXTUALIZED KNOWLEDGE

A number of social cognitive studies have shown that the integration of information across communicative modalities can facilitate the language comprehension of human readers (Jones and LeBaron, 2002; Calero, 2005; Brinke and Weisbuch, 2020). In this chapter, we focus on integrating verbal (e.g., utterances of speakers) and nonverbal information (e.g., body movements, facial expressions, vocal tones, or mental states of speakers) originally conveyed in different modalities within a short time period and exploring the influence of verbal-nonverbal knowledge on machine reading comprehension tasks, especially those non-extractive MRC tasks studied in the previous chapters that contain a high proportion of questions requiring general world knowledge unstated in the given documents (Mostafazadeh et al., 2016; Lai et al., 2017; Ostermann et al., 2018; Sun et al., 2019a; Huang et al., 2019a; Sun et al., 2020b).

Typically, each piece of structured knowledge is represented as a triple that contains two phrases (e.g., (*"finding a lost item"*, *"happiness"*) and the relation (e.g., CAUSES) between phrases, which can be one of a pre-defined set of relations (Tandon et al., 2014; Speer et al., 2017; Grishman, 2019). A carefully designed relation set is indispensable for many fundamental tasks such as knowledge graph construction. However, it is still unclear whether we need to explicitly represent relations if the final goal is to help downstream tasks that do not directly depend on the reliability of relations in triples from external sources. Once we decide not to name relations, one natural question is whether we could implicitly represent relations between two phrases. We suggest that adding context in which the

phrases occur may be helpful as such a context constrains the possible relations between phrases without intervening in the relations explicitly (Brézillon et al., 1998). Hereafter, we call a triple that contains a phrase pair and its associated context as a piece of **contextualized knowledge**.

To represent verbal-nonverbal knowledge, we regard related verbal and nonverbal information as the phrase pair; we treat the text in which the verbal-nonverbal pair occurs as the context. We suggest film and television show scripts are good source corpora for extracting contextualized verbal-nonverbal knowledge as they contain rich strongly interrelated verbal and nonverbal information, which can be easily separated from the scripts. For example, as shown in Table 6.1, the pause in *"I'm going.....to his house."* is related to *"thinking"*, the internal state of the speaker. Furthermore, a script usually contains multiple scenes, and the entire text of the scene from which the verbal-nonverbal pair is extracted can serve as the context. According to the relative position of a verbal-nonverbal pair in a scene, we use lexical patterns to extract four types of contextualized knowledge (Section 6.1).

To make the form of knowledge suitable for MRC tasks, we randomly select nonverbal messages from the same script to convert each piece of knowledge into a weakly-labeled MRC instance (Section 6.2). We propose a two-stage fine-tuning strategy to use the weakly-labeled MRC data: first, we train a model on the combination of the weakly-labeled data and the target MRC data that is human-annotated but relatively small-scale, and then, we fine-tune the resulting model on the target data alone (Section 6.3). We observe that training over the combination of all the data based on all types of contextualized knowledge does not lead to noticeable gains compared to using one type of knowledge.

Therefore, we further adopt a teacher-student paradigm with multiple teacher models trained with different types of knowledge (Section 6.4).

As a starting point, we focus on scripts written in Chinese. Hence, we evaluate our method on C^3 presented in the previous chapter, as far as we know, the only multiple-choice MRC dataset for Chinese wherein most questions require general world knowledge beyond the given contents. Experimental results demonstrate that our method leads to +4.3% in accuracy over a state-of-the-art baseline (Cui et al., 2020). We also seek to transfer the knowledge to help other tasks by adapting the resulting student MRC model, yielding gains in both noisy and clean settings: up to +7.8% in accuracy on MRC datasets DREAM (Chapter 4) and Cosmos QA (Huang et al., 2019a), which are automatically translated from English into Chinese, and +2.9% in F1 on the Chinese set of a bilingual document-level relation extraction dataset DialogRE (Yu et al., 2020) over strong baselines. These results indicate the usefulness of the extracted knowledge.

This chapter is based on Sun et al. (2020a).

6.1 Contextualized Knowledge Extraction

We regard that understanding the interactions between verbal and nonverbal messages may require general world knowledge as they function together in communications, and such knowledge is assumed to be known by most people without being taught. We propose to use interrelated verbal and nonverbal information as phrases in the classical triple-style knowledge representation and situate them in a context. Formally, we call a triple (v, c, n) as a piece of **contextualized knowledge**, containing a pair of related verbal information v

Scene 1	
□	Interior. Runaway office. Day.
Andy:	I tried to ask her, but...
Emily:	You never ask Miranda. Anything. (**sighs**) All right, I'll take care of the other stuff. You go to Calvin Klein.
Andy:	Me?
Emily:	I'm sorry. Do you have a prior commitment? Is there some hideous pants convention?
Andy:	So I just, what, go down to the Calvin Klein store and ask them...
◊	**Emily rolls her eyes so hard they almost eject from her head.**
Emily:	You're not going to the store.
Andy:	Of course not. I'm going...(**thinking**)...to his house.
Emily (**oh god**):	You are catching on quickly. We always send assistants to a designer's home on their very first day. You're going to his showroom. I'll give you the address.
Andy:	Sorry. Got it. What's the nearest subway stop?
Emily:	Good God. You do not. Under any circumstances. Take public transportation.
Andy:	I don't?

type	**nonverbal**	**verbal**
B_c	oh god	Emily: You are catching on [...] I'll give you the address.
I	sighs	Emily: You never ask Miranda. Anything. All right [...] Klein.
I	thinking	Andy: Of course not. I'm going.....to his house.
O	Emily rolls her eyes so hard they almost eject from her head.	Andy: So I just, what, go down to the Calvin Klein store and ask them...

Table 6.1: A sample scene in a script and examples of extracted verbal-nonverbal pairs from this scene (all translated into English; [...]: words omitted; □: scene heading; ◊: action line). The scene is regarded as the context of all the verbal-nonverbal pairs.

and nonverbal information n, and the associated context c. We choose to extract contextualized knowledge from film and television show scripts[1] as plentiful verbal and nonverbal messages frequently co-occur in scripts, and they can be easily separated. Scenes in a script are separated by blank lines. According to the relative position of verbal and nonverbal information, we extract four types of contextualized knowledge (B_c, B_n, I, and O):

- Beginning: the nonverbal information n appears after a speaker name and before the speaker's utterance. We regard the speaker name and the

[1]As it is difficult to verify whether a text is written before a presentation (i.e., script) or during/after a presentation (i.e., transcript), we use *scripts* throughout this chapter.

corresponding utterance as *v*.

- Clean (B_c): We only extract nonverbal information *n* within parentheses.
- Noisy (B_n): The first span of a turn, followed by a colon, can also contain both a speaker name and nonverbal information about this speaker (e.g., *"Xiaocong Le took the cup of hot water: 'Thank you!' "*). We remove the phrase that is a potential speaker name from the span and regard the remaining text in the span as *n*. We roughly regard a phrase as a speaker name if it appears in the first span of other turns in the same scene.

- Inside (I): We only extract nonverbal information *n* enclosed in parentheses, which appears within an utterance. All the information in the same turn except *n* is treated as *v*.
- Outside (O): Here *n* is an action line that mainly describes what can be seen or heard by the audience, marked by ◊ in Table 6.1. We regard the turn (if it exists) before the action line as its corresponding *v*.

We do not extract phrases in parentheses or action lines as nonverbal information if they are terminologies for script writing such as *"O.S."*, *"V.O."* *"CONT'D"*, *"beat"*, *"jump cut"*, and *"fade in"*. All types of contextualized knowledge extracted from a scene share the same context *c*, i.e., the scene itself. We do not exploit the scene heading mostly about when and where a scene takes place (marked by □ in Table 6.1), as it is intentionally designed to cover the content of the whole scene, which is already used as context. See more extracted contextualized verbal-nonverbal knowledge triples in Table 6.2 and Table 6.3.

type	nonverbal	verbal
Scene 2		
Laifu Zhao:	Le Xiaocong, it is you! I thought there was a thief in the kitchen.	
Xiaocong Le:	Monitor, I got in without your permission. I'm really sorry.	
Laifu Zhao:	It's okay, it's okay. He handed a cup of hot water:	
	This room is too cold. Come here and take a sip of hot water to warm up. Take a break.	
Xiaocong Le took the cup of hot water:	Thank you! Monitor!	
Laifu Zhao:	Hey, Le Xiaocong, see how dedicated and focused you are. What are you writing?	
B_n	took the cup of hot water	Xiaocong Le: Thank you! Monitor!
Scene 3		
Wukong:	Huh! Why does this picture move?	
Aoguang:	This is a cruiser. The picture on the screen is a submarine fiber optic cable being laid.	
Wukong:	A cruiser?	
Aoguang:	Yes. With it, finding a needle in a haystack is not difficult. Hahaha...	
Wukong:	Finding Huaguoshan is easier than finding a needle in a haystack!	
Aoguang dumbfounded:	Well, well, well...	
B_n	dumbfounded	Aoguang: Well, well, well...
Scene 4		
	From a cauldron on the stove, hot water is poured into two pails, by the a kitchen boy under the nurse's command.	
Scullery Maid (O.S.):	Thomas Kent, sir? No sir.	
Will (O.S.):	The actor.	
Nurse:	Who asks for him?	
	Will has come to the kitchen door with a letter.	
Will:	William Shakespeare, actor, poet, and playwright of the Rose.	
	The nurse sends the scullery maid back to work.	
Nurse:	Master Kent is. . . my nephew.	
Will (giving her the letter):	I will wait.	
Nurse:	Much god may it do you.	
B_c	giving her the letter	Will: I will wait.
O	Will has come to the kitchen door with a letter.	Nurse: Who asks for him?
O	The nurse sends the scullery maid back to work.	Will: William Shakespeare, actor, poet, and playwright of the Rose.

Table 6.2: Additional examples of extracted verbal-nonverbal pairs situated in scenes (Part 1, all translated into English).

type	nonverbal	verbal
Scene 5		
	The ball is lowered, lighting up a sign that reads "1972."	
Crowd(over television):	...1! Happy New Year!	
	The people in the bar cheer and kiss each other. They blow horns and toss confetti into the air.	
	Forrest looks around as Carla and Lenore lean over and kiss him.	
B_c	over television	Crowd: ...1! Happy New Year!
Scene 6		
	Scene 90: outside the field	
Principal Qi:	(nodding his head) Excellent Game! Good job! Give students some water!	
Dean of Education:	Water is here! (distributed bottles of water one by one)	
Haiyin Lin:	Everyone played well in today's game!	
Gao Feng:	Unfortunately, we still lost by two points.	
	Several people nodded	
I	nodding his head	Principal Qi: Excellent Game! Good job! Give students some water!
I	distributed bottles of water one by one	Dean of Education: Water is here!
O	Several people nodded.	Gao Feng: Unfortunately, we still lost by two points.
Scene 7		
Mrs. Gump:	You do your very best now, Forrest.	
Forrest:	I sure will, Momma.	
Forrest (V.O.):	I remember the bus ride on the first day of school very well.	
Bus Driver:	Are you comin' along? Forrest: Momma said not to be taking rides from strangers.	
Bus Driver:	This is the bus to school.	
Forrest:	I'm Forrest Gump.	
Bus Driver:	I'm Dorothy Harris.	
Forrest:	Well, now we ain't strangers anymore.	
	The bus driver smiles as Forrest steps up into the bus.	
O	The bus driver smiles as Forrest steps up into the bus.	Forrest: Well, now we ain't strangers anymore.

Table 6.3: Additional examples of extracted verbal-nonverbal pairs situated in scenes (Part 2, all translated into English).

6.2 Instance Generation

As most current MRC tasks requiring general world knowledge are usually in a multiple-choice form, we mainly discuss how to convert the extracted triples into multiple-choice instances and leave its extension to other types (e.g., extractive or abstractive) of MRC tasks for future research. We generate instances for each type of contextualized knowledge. For each triple (v, c, n), we remove n from context c, and we regard the remaining content as the reference document, verbal information v as the question, and the nonverbal information n as the correct answer option. To generate distractors (i.e., wrong answer options), we randomly select N items from all the unique nonverbal information in other triples, which belong to the same type of contextualized knowledge and are extracted from the same script as (v, c, n). Note that we only generate one instance based on each triple, while it is easy to generate more instances by changing distractors.

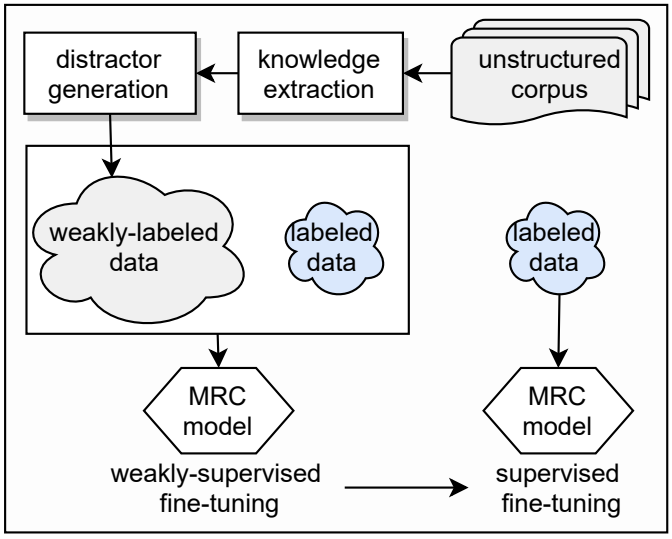


Figure 6.1: Two-stage fine-tuning framework overview (one type of contextualized knowledge is involved).

6.3 Two-Stage Fine-Tuning

As mentioned previously, we aim to use the constructed weakly-labeled data to improve a downstream MRC task. Given weakly-labeled data generated based on **one** type of contextualized knowledge (e.g., B_c or I) extracted from scripts, we first use the weakly-labeled data in conjunction with the training set of the target MRC data as the training data to train the model and then fine-tune the resulting model on the target MRC data as illustrated in Figure 6.1. We do not adjust the ratio of clean data to weakly-labeled data observed during training as previous joint training work on other tasks such as machine translation (Edunov et al., 2018).

Another way is to perform separate training: we first train the model on the weakly-labeled data and then fine-tune it on the target data. In our preliminary experiment, we observe that joint training leads to better performance, and hence we apply it in all the experiments. See performance comparisons of joint and separate training in Section 6.5.4.

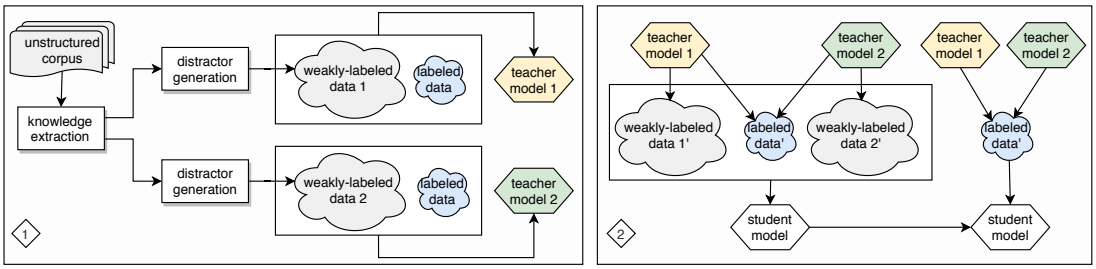


Figure 6.2: Teacher-student paradigm overview (multiple types of contextualized knowledge are involved). To save space, we only show the case that involves two types of contextualized knowledge.

6.4 Teacher-Student Paradigm

As introduced in Section 6.2, we have **multiple** sets of weakly-labeled data, each corresponding to one type of contextualized knowledge (Section 6.1). We observe that simply combining all the data, either in joint training or separate training, does not lead to noticeable gains compared to using one type of contextualized knowledge. Inspired by the previous work (You et al., 2019) that trains a student automatic speech recognition model with multiple teacher models, and each teacher model is trained on a domain-specific subset with a unique speaking style, we employ a teacher-student paradigm to inject multiple types of contextualized knowledge into a single student machine reader.

Let V denote a set of labeled instances, $W_1, \ldots, W_\ell$ denote ℓ sets of weakly-labeled instances, and $W = \bigcup_{1 \le i \le \ell} W_i$. For each instance t, we let m_t denote its total number of answer options, and $\boldsymbol{h}^{(t)}$ be a hard label vector (one-hot) such that $\boldsymbol{h}^{(t)}_j = 1$ if the j-th option is labeled as correct. We train ℓ teacher models, denoted by $\mathcal{T}_1, \ldots, \mathcal{T}_\ell$, and optimize $\mathcal{T}_i$ by minimizing $\sum_{t \in V \cup W_i} L_1(t, \theta_{\mathcal{T}_i})$. L_1 is defined as

$$L_1(t, \theta) = -\sum_{1 \le k \le m_t} \boldsymbol{h}^{(t)}_k \log p_\theta(k \mid t),$$

where $p_\theta(k \mid t)$ denotes the probability that the k-th option of instance t is correct, estimated by the model with parameters θ.

We define soft label vector $\boldsymbol{s}^{(t)}$ such that

$$\boldsymbol{s}^{(t)}_k = \begin{cases} \lambda\, \boldsymbol{h}^{(t)}_k + (1-\lambda) \displaystyle\sum_{1 \le j \le \ell} \frac{1}{\ell} p_{\theta_{\mathcal{T}_j}}(k \mid t) & t \in V \\ \lambda\, \boldsymbol{h}^{(t)}_k + (1-\lambda) p_{\theta_{\mathcal{T}_i}}(k \mid t) & t \in W_i \end{cases},$$

where $\lambda \in [0, 1]$ is a weight parameter, and $k = 1, \ldots, m_t$.

We then train a student model, denoted by $\mathcal{S}$, still in a two-stage fashion. In

stage one (i.e., weakly-supervised fine-tuning), we optimize $\mathcal{S}$ by minimizing $\sum_{t \in V \cup W} L_2(t, \theta_{\mathcal{S}})$, where L_2 is defined as

$$L_2(t, \theta) = -\sum_{1 \le k \le m_t} \boldsymbol{s}_k^{(t)} \log p_\theta(k \mid t).$$

In stage two (i.e., supervised fine-tuning), we further fine-tune the resulting $\mathcal{S}$ after stage one by minimizing $\sum_{t \in V} L_2(t, \theta_{\mathcal{S}})$. See Figure 6.2 for an overview of the paradigm.

6.5 Experiment

6.5.1 Data

We collect 8,166 scripts in Chinese, and most of them are intended for films and television shows.[2] After segmentation and filtering, we obtain 199,280 scenes, each of which contains at least one piece of contextualized knowledge defined in Section 6.1. We generate four sets of weakly-labeled MRC data based on four types of contextualized knowledge. For comparison, we also use existing human-annotated triples about general world knowledge in the Chinese version of ConceptNet (Speer et al., 2017). We set the number of distractors N (Section 6.2) to five for weakly-labeled MRC instances.

For evaluation, we use C^3, so far as we know, the only multiple-choice MRC data for Chinese with a focus on general world knowledge. About 86.8% of questions in C^3 involve prior knowledge (i.e., linguistic, domain-specific, and general world knowledge) unstated in the given texts, and all instances are

[2] `https://www.1bianju.com`.

carefully designed by experts such as second-language teachers. Each instance consists of a document, a question, and multiple answer options; only one option is correct. Furthermore, we use Google Translate to generate Chinese versions of DREAM and Cosmos QA, two popular multiple-choice MRC datasets for English in which most questions require general world knowledge, as additional indications to evaluate the usefulness of the extracted knowledge. Besides MRC tasks, we use the Chinese set of a bilingual relation extraction dataset DialogRE, which also requires document-level understanding to predict relations from 36 possible types between an argument pair. See Table 6.4 for data statistics. While we focus mostly on resources in Chinese, our extraction and training methods are not limited to a particular language.

data	type of construction	# of instances
C^3	human-annotated	19,577
DREAM	human-annotated	10,197
Cosmos QA	human-annotated	35,600
DialogRE	human-annotated	10,886
ConceptNet	human-annotated	737,534
B_c	weakly-labeled	105,622
B_n	weakly-labeled	198,053
I	weakly-labeled	204,750
O	weakly-labeled	192,391
B_c+ B_n+ I + O	weakly-labeled	700,816

Table 6.4: Data Statistics.

6.5.2 Implementation Details

We follow Sun et al. (2020b) for the model architecture consisting of a pre-trained language model and a classification layer on top of the model. We use RoBERTa-wwm-ext-large (Cui et al., 2020) as the pre-trained language model, which achieves state-of-the-art performance on C^3 and many other natural lan-

index	weakly-supervised fine-tuning		supervised fine-tuning		dev	test
	data	teacher-student	data	teacher-student		
0	–	–	C^3	–	73.9	73.4
1	C^3+ B_c	–	–	–	71.1	71.7
2	C^3+ B_c	–	C^3	–	74.5	74.0
3	C^3+ B_n	–	–	–	71.3	72.0
4	C^3+ B_n	–	C^3	–	74.6	74.5
5	C^3+ I	–	–	–	73.5	72.8
6	C^3+ I	–	C^3	–	**75.6**	**74.9**
7	C^3+ O	–	–	–	72.4	72.7
8	C^3+ O	–	C^3	–	75.4	74.9
9	C^3+ B_c+ B_n+ I + O	–	–	–	71.6	71.0
10	C^3+ B_c+ B_n+ I + O	–	C^3	–	75.6	75.2
11	C^3+ B_c+ B_n+ I + O	✓	C^3	–	76.5	76.4
12	C^3+ B_c+ B_n+ I + O	✓	C^3	✓	**77.4**	**77.7**

Table 6.5: Average accuracy (%) on the development and test sets of the C^3 dataset.

guage understanding tasks in Chinese (Xu et al., 2020). We leave the exploration of more pre-trained language models for future work. When the input sequence length exceeds the limit, we repeatedly discard the last turn in the context, or the first turn if the last turn includes the extracted verbal information. We train a model for one epoch during the weakly-supervised fine-tuning stage and eight epochs during the supervised fine-tuning stage. We set λ (defined in Section 6.4) to 0.5 in all experiments based on the rationale that we can make best use of the soft labels while at the same time making sure $\arg\max_k \boldsymbol{s}_k^{(t)}$ is always the index of the correct option for instance t. Carefully tuning λ on the development set may lead to further improvements, which is not the primary focus of this chapter.

6.5.3 Main Results and Discussions

Table 6.5 reports the main results. The baseline accuracy (73.4% {0}) is slightly lower than previously reported using the same language model[3] as we report the average accuracy over five runs with different random seeds for all our supervised fine-tuning results. For easy reference, we indicate the index for each result in curly brackets in the following discussion. Obviously, the performance of a model after the first fine-tuning stage over the combination of the C^3 dataset and much larger weakly-labeled data is worse (e.g., 71.7% {1}) than baseline performance ({0}). Further fine-tuning the resulting model on the C^3 dataset consistently leads to improvements (e.g., 74.0% {2} and 74.5% {4}) over the baseline {0}, demonstrating the effectiveness of the **two-stage** fine-tuning strategy for using large-scale weakly-labeled data. We will discuss the critical role of the target task's data (i.e., C^3) in the weakly-supervised fine-tuning stage in the next subsection. Following this strategy, **each** of the weakly-labeled data based on one type of contextualized knowledge can boost the final performance ({2, 4, 6, 8}); the magnitude of accuracy improvement is 1.2% on average.

When we combine all the weakly-labeled data in the first fine-tuning stage, the performance gain after the second round of fine-tuning (75.2% {10}) is not as impressive as expected, given the best performance achieved by only using one set (74.9% {6}). As a comparison, our **teacher-student paradigm** that trains multiple teacher models with different types of weakly-labeled data leads to up to 3.7% improvement in accuracy ({12} vs. {2, 4, 6, 8}). The advantage is reduced but still exists even when we use the original hard labels instead of soft labels in the second fine-tuning stage (76.4% {11}).

[3] https://github.com/CLUEbenchmark/CLUE.

6.5.4 Ablation Studies and Analysis

We have shown that the present teacher-student paradigm helps inject multiple types of knowledge into a reader. We conduct two ablation studies to examine critical factors. We remove the context (i.e., scene) from each instance in the weakly-labeled data and leave it empty. All other aspects of this baseline remain the same as {12} in Table 6.5. We also experiment with removing C^3 from the weakly-supervised fine-tuning stage when we train teacher and student models (Figure 6.2) for comparisons. We observe that accuracy decreases in both cases (Table 6.6), demonstrating the usefulness of contexts in contextualized knowledge for improving MRC and the importance of involving the human-annotated data of the target task, although small-scale, in the weakly-supervised fine-tuning stage.

method	dev	test
{12} in Table 6.5	77.4	77.7
{12} w/o context in weakly-labeled data	76.8	76.6
{12} w/o using C^3 in the 1st FT	76.6	76.2

Table 6.6: Ablation results on the development and test sets of the C^3 dataset (FT: fine-tuning).

It is difficult, however, to infer which pieces of knowledge help the improved MRC instances. As an alternative solution, we study the impacts of the contextualized knowledge on different types of questions based on the annotated subset (300 instances) released along with the dataset. As shown in Table 6.7, our method generally improves performance on all types of questions, especially those that require general world knowledge. In particular, we observe accuracy improvements of 10.0% or more on questions that require cause-effect, part-whole, or scenario, three subcategories of general world knowledge. For instance, given a conversation, *"Female: Sir, can you drive faster? I'm afraid that I*

will be late for the exam. Male: No, the speed is already quite fast. Safety is also very important.", we require scenario knowledge about activities of humans, their corresponding location information, and personal information such as the profession, in order to answer the question about the possible location (*"taxi"*) of the two speakers.

category	{0}	{12}	Δ
Matching	90.0	**94.7**	4.7
Prior Knowledge	69.5	**75.3**	5.8
··· Linguistic	73.8	**77.8**	4.0
··· General world knowledge	68.0	**74.4**	6.4
··· Arithmetic	34.3	**40.0**	5.7
··· Connotation	74.0	**78.0**	4.0
··· Cause-effect	78.0	**88.0**	10.0
··· Implication	68.5	**70.8**	2.3
··· Part-whole	58.2	**70.9**	12.7
··· Precondition	60.0	**65.7**	5.7
··· Scenario	64.8	**76.5**	11.7
··· Domain-specific⋆	13.3	**20.0**	6.7

Table 6.7: Average accuracy (%) on the annotated development set of C^3 per category (⋆: only three instances).

notes	weakly-labeled data				dev	test
	structured knowledge	document	question	answer		
{0} in Table 6.5	–	–	–	–	73.9	73.4
{10} in Table 6.5	contextualized knowledge	scene	verbal	nonverbal	75.6	75.2
{10} w/o context	contextualized knowledge	empty	verbal	nonverbal	74.9	74.2
i	ConceptNet	empty	subject	object	74.0	72.7
ii	ConceptNet	relation type	subject	object	74.6	74.1

Table 6.8: Average accuracy (%) on the development and test sets of the C^3 dataset using weakly-labeled data constructed based on contextualized knowledge or ConceptNet.

6.5.5 A Comparison Between Contextualized Knowledge and ConceptNet

From the perspective of improving a downstream MRC task, we compare the extracted contextualized knowledge with standard general world knowledge graphs, which have been shown to improve MRC tasks (Wang et al., 2018c). As most of such graphs are in English, we only compare contextualized knowledge with the human-annotated Chinese version of ConceptNet. Each triple in ConceptNet is represented as (subject, relation type, object) (e.g., (*"wing"*, PART_OF, *"an airplane"*)). We experiment with two types of input sequences when we convert triples into MRC instances: (i) leave the document empty in each instance and (ii) use the relation type as the document. We randomly select phrases in ConceptNet other than the phrases in each triple as distractors.

For a fair comparison, we compare (ii) with baseline {10} in Table 6.5 as it follows the same two-stage fine-tuning without using the teacher-student paradigm. To compare with (i), we run an ablation test of {10} by removing contexts from weakly-labeled MRC instances. The amounts of weakly-labeled instances based on contextualized knowledge and ConceptNet are similar (Table 6.4). The results in Table 6.8 reveal that under the two-stage fine-tuning framework, introducing ConceptNet yields up to 0.7% in accuracy, but using contextualized knowledge gives a bigger gain of 1.8% in accuracy. Moreover, removing contexts from weakly-labeled instances hurts performance, consistent with our observation in Section 6.5.4.

We admit that our knowledge representation is not concise enough for easy alignment with existing graphs. Nevertheless, we argue that contexts can tacitly

state relations between phrases and emphasize the usefulness of contextualized knowledge for MRC tasks requiring general world knowledge.

6.5.6 The Usefulness of Contextualized Knowledge for Other Tasks

We report the average accuracy over five runs with different random seeds for all results. For MRC datasets DREAM and Cosmos QA, which are in translated Chinese, we simply use {12} in Table 6.5 to initialize an MRC model. As shown in Table 6.9, in this noisy setting, we still obtain 7.8% in accuracy on the test set of DREAM and 2.7% in accuracy on the publicly available development set of Cosmos QA, by adapting our best-performing MRC model. The different performance levels on translated datasets and their original English versions may be due to the different sizes of text corpora for pre-training language models for English and Chinese and noise introduced by imperfect automatic machine translation.

parameter initialization	DREAM		Cosmos QA
	dev	test	dev
RoBERTa-wwm-ext-large	61.4	60.8	56.7
{0} in Table 6.5	67.0	65.5	57.1
{12} in Table 6.5	**69.2**	**68.6**	**59.4**

Table 6.9: Average accuracy on the translated Chinese version of DREAM and Cosmos QA.

For DialogRE, instead of converting the extracted triples into weakly-labeled relation extraction instances and training from scratch, we simply replace the classification layer of an MRC model with a multi-class multi-label classification layer following the baseline released by Yu et al. (2020) and fine-tune the whole

parameter initialization	dev		test	
	F1	$F1_c$	F1	$F1_c$
$BERT_S$ (Yu et al., 2020)	65.5	61.0	63.5	58.7
RoBERTa-wwm-ext-large	64.9	60.3	64.4	59.2
{0} in Table 6.5	66.4	61.6	65.0	60.3
{12} in Table 6.5	**67.1**	**62.9**	**67.3**	**62.3**

Table 6.10: Average F1 (%) and $F1_c$ (%) on DialogRE.

architecture on DialogRE. We compare the performance of methods that use different weights for parameter initialization except for the randomly initialized classification layer. We achieve +2.9% in F1 and +3.1% in $F1_c$ on DialogRE (Table 6.10). The metric $F1_c$ is used to encourage a model to identify relations between arguments as early as possible rather than after reading the whole dialogue. Introducing C^3 alone also allows us to achieve a slight gain over the relation extraction baseline. It might be interesting to investigate the relevance between document-level relation extraction and machine reading comprehension for further performance boost.

6.6 Related Work

6.6.1 Contextualized Knowledge Extraction

Here we primarily discuss external contextualized knowledge that is not directly relevant with a target task as retrieving related pieces of evidence from an external source for each instances of a target task is not our focus. A common solution to obtain external contextualized knowledge is to utilize existing knowledge bases via distant supervision (Ye et al., 2019). We extract contextualized knowledge from scripts, wherein contexts (i.e., scenes) are naturally aligned with

verbal-nonverbal pairs to avoid noise. Besides, we focus on improving MRC with verbal-nonverbal knowledge, which is seldom studied.

Our work is also related to commonsense knowledge extraction, which relies on human-annotated triples (Xu et al., 2018a; Bosselut et al., 2019), high-precision syntactic or semantic patterns (Zhang et al., 2020; Zhou et al., 2020) specific to each relation, or existing lexical databases (Tandon et al., 2014, 2015). By contrast, we skip the step of offering a name of the relation between phrases and situate structured knowledge in its context. Our language-independent knowledge extraction does not require any training data and does not rely on a high-quality semantic lexicon or a syntactic parser, which is not always available.

6.6.2 Weak Supervision and Semi-Supervised Learning for MRC

As it is expensive and time-consuming to crowdsource or collect a large-scale, high-quality dataset, weak supervision has received much attention throughout the MRC literature. Various forms of weak supervision are studied, mostly based on existing resources such as pre-trained semantic/syntactic parsers (Smith et al., 2015; Wang et al., 2015; Liu et al., 2017) or natural language inference systems (Pujari and Goldwasser, 2019; Wang et al., 2019), knowledge bases (Wang and Jiang, 2019; Yang et al., 2019), and linguistic lexicons (Sun et al., 2019c). Compared to previous work, we focus on generating large-scale weakly-labeled data using the contextualized knowledge automatically extracted from unstructured corpora.

Previous semi-supervised methods that leverage internal or external unla-

beled texts usually generate question and answer based on the content of the same sentence (Yang et al., 2017; Wang et al., 2018b; Dhingra et al., 2018). Besides the unlabeled texts, previous studies (Yuan et al., 2017; Yu et al., 2018; Zhang and Bansal, 2019; Zhu et al., 2019; Dong et al., 2019; Alberti et al., 2019; Asai and Hajishirzi, 2020) also heavily rely on the labeled instances of the target MRC task for data augmentation. In comparison, we generate weakly-labeled MRC instances without using any task-specific patterns or labeled data to improve MRC tasks that require substantial general world knowledge. Another line of work develops unsupervised approaches (Lewis et al., 2019; Li et al., 2020; Fabbri et al., 2020) for extractive MRC tasks. However, there is still a large performance gap between unsupervised and state-of-the-art supervised methods.

6.6.3 Knowledge Utilization

Our teacher-student paradigm for knowledge utilization is most related to multi-domain teacher-student training for automatic speech recognition (You et al., 2019) and machine translation (Wang et al., 2020). Instead of clean domain-specific human-labeled data, each of our teacher models is trained with **weakly-labeled data**. Due to the introduction of large amounts of weakly-labeled data, the data of the target MRC task (with hard or soft labels) is used during all the fine-tuning stages of both teacher and student models.

6.7 Chapter Summary

This chapter introduces how to extract contextualized verbal-nonverbal knowledge from film/TV scripts and use this kind of knowledge to improve machine reading comprehension. We propose to situate structured knowledge in a context to implicitly represent the relations between phrases, instead of relying on a predefined set of relations. We propose a two-stage fine-tuning strategy to use the large-scale weakly-labeled data and employ a teacher-student paradigm to inject multiple types of contextualized knowledge into a single student model. Experimental results show that our method outperforms a state-of-the-art baseline by +4.3% in accuracy on the multiple-choice MRC dataset C^3. Finally, we show the usefulness of the extracted knowledge for other MRC task and MRC-related tasks such as document-level relation extraction.

CHAPTER 7

CONCLUSION AND FUTURE WORK

In this dissertation, we present our work in developing approaches to improve non-extractive MRC and the creation of datasets that pose new challenges for MRC systems. In this chapter, we summarize the contributions of this dissertation and discuss directions for future work.

7.1 Summary of Contributions

In Chapter 3, we propose three general strategies to improve non-extractive machine reading comprehension: BACK AND FORTH READING, HIGHLIGHTING, and SELF-ASSESSMENT. By fine-tuning a pre-trained language model with our proposed strategies on the largest general domain multiple-choice MRC dataset RACE, we obtain a 5.8% absolute increase in accuracy over the previous best result achieved by the same pre-trained model fine-tuned on RACE without the use of strategies. We further fine-tune the resulting model on a target MRC task, leading to an absolute improvement of 6.2% in average accuracy over previous state-of-the-art approaches on six representative non-extractive MRC datasets from different domains. These results demonstrate the effectiveness of our proposed strategies and the versatility and general applicability of our fine-tuned models that incorporate these strategies.

In Chapter 4, we present DREAM, the first dialogue-based multiple-choice reading comprehension dataset. Collected from English-as-a-foreign-language examinations designed by human experts to evaluate the comprehension level of

Chinese learners of English, our dataset contains 10,197 multiple-choice questions for 6,444 dialogues. DREAM is the first dataset to focus on in-depth multi-turn multi-party dialogue understanding and is likely to present significant challenges for reading comprehension systems: 84% of answers are non-extractive, 85% of questions require reasoning beyond a single sentence, and 34% of questions also involve commonsense knowledge. We apply several popular neural reading comprehension models that primarily exploit surface information within the text and find them to, at best, just barely outperform a rule-based approach. We next investigate the effects of incorporating dialogue structure and different kinds of general world knowledge into both rule-based and (neural and non-neural) machine learning-based reading comprehension models. Experimental results on the DREAM dataset show the effectiveness of dialogue structure and general world knowledge.

In Chapter 5, we present the first free-form multiple-Choice Chinese machine reading Comprehension dataset (C^3), 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 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 and human readers, 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 C^3 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 C^3 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.

In Chapter 6, we develop a method of utilizing contextualized verbal-nonverbal knowledge extracted from film/TV scripts to improve machine reading comprehension tasks that require tacit general world knowledge. Experimental results show that our method outperforms a state-of-the-art baseline by +4.3% in accuracy on C^3. We also seek to transfer the knowledge to other tasks by simply adapting the resulting model, yielding up to +7.8% in accuracy on translated MRC datasets such as DREAM and Cosmos QA and +2.9% in F1 on a relation extraction dataset DialogRE that also involves document-level reading comprehension, demonstrating the usefulness of contextualized verbal-nonverbal knowledge for MRC.

7.2 Future Work

Downstream application. Compared with extractive MRC, fewer works have been dedicated to the practical application of non-extractive multiple-choice MRC. Besides formulating a target problem as an MRC task and directly applying MRC techniques, we can also explore ways of leveraging the intermediate representations of MRC models when tackling problems that require language understanding (e.g., in Chapter 6, we show that we can transfer knowledge from a pre-trained MRC model to a relation extraction model) to apply MRC techniques to a broader range of problems.

Explainability in machine meading comprehension. As the MRC research direction evolves from comprehending explicitly expressed information to in-depth understanding that requires advanced reading skills and prior world knowledge, it becomes more challenging and critical than ever to develop techniques to enable MRC models to deliver the answer explanation, on top of the ability to arrive at the answer.

Long-text reading comprehension. The documents in most MRC research, including the proposed and studied tasks in this dissertation, are short, typically no longer than a few hundred words. Ultimately, we would like to create reading comprehension systems that can achieve genuine human-level reading comprehension performance, and one of the desired abilities is the ability of reading long documents such as books. There is still a long way to go towards long-text reading comprehension, as it is even unrealistic for most state-of-the-art MRC systems to encode a book-length document.

BIBLIOGRAPHY

Marilyn Adams and Bertram Bruce. 1982. Background knowledge and reading comprehension. *Reader meets author: Bridging the gap*, 13:2–25.

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the ACL*, pages 6168–6173.

Leila Amgoud, Yannis Dimopoulos, and Pavlos Moraitis. 2007. A unified and general framework for argumentation-based negotiation. In *Proceedings of the AAMAS*, pages 1–8, New York, NY, USA.

Akari Asai and Hannaneh Hajishirzi. 2020. Logic-guided data augmentation and regularization for consistent question answering. In *Proceedings of the ACL*, pages 5642–5650.

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

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. 2014. Modeling biological processes for reading comprehension. In *Proceedings of the EMNLP*, pages 1499–1510, Doha, Qatar.

Steven Bird and Edward Loper. 2004. NLTK: the natural language toolkit. In *Proceedings of the ACL on Interactive poster and demonstration sessions*, pages 31–34, Barcelona, Spain.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for

automatic knowledge graph construction. In *Proceedings of the ACL*, pages 4762–4779.

Florian Boudin, Hugo Mougard, and Benoit Favre. 2015. Concept-based summarization using integer linear programming: From concept pruning to multiple optimal solutions. In *Proceedings of the EMNLP*, pages 17–21, Lisbon, Portugal.

Patrick Brézillon, J-Ch Pomerol, and Ilham Saker. 1998. Contextual and contextualized knowledge: An application in subway control. *International Journal of Human-Computer Studies*, 48(3):357–373.

Leanne Brinke and Max Weisbuch. 2020. How verbal-nonverbal consistency shapes the truth. *Journal of Experimental Social Psychology*, 89.

Henry H Calero. 2005. *The power of nonverbal communication: How you act is more important than what you say*. Silver Lake Publishing.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of the ACL*, pages 789–797, Columbus, OH.

Eugene Charniak. 1972. *Toward a model of children's story comprehension*. Ph.D. thesis, Massachusetts Institute of Technology.

Danqi Chen. 2018. *Neural Reading Comprehension and Beyond*. Ph.D. thesis, Stanford University.

Danqi Chen, Jason Bolton, and Christopher D Manning. 2016. A thorough examination of the CNN/Daily Mail reading comprehension task. In *Proceedings of the ACL*, pages 2358–2367, Berlin, Germany.

Yu-Hsin Chen and Jinho D Choi. 2016. Character identification on multiparty

conversation: Identifying mentions of characters in TV shows. In *Proceedings of the SIGDial*, pages 90–100, Los Angeles, CA.

Zhipeng Chen, Yiming Cui, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. Convolutional spatial attention model for reading comprehension with multiple-choice questions. In *Proceedings of the AAAI*, Honolulu, HI.

Zhipeng Chen, Yiming Cui, Wentao Ma, Shijin Wang, Ting Liu, and Guoping Hu. 2018. HFL-RC system at SemEval-2018 Task 11: Hybrid multi-aspects model for commonsense reading comprehension. *arXiv preprint*, cs.CL/1803.05655v1.

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 City, 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.

Yu-An Chung, Hung-yi Lee, and James Glass. 2018. Supervised and unsupervised transfer learning for question answering. In *Proceedings of the NAACL-HLT*, pages 1585–1594, New Orlean, LA.

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, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint*, cs.CL/1803.05457v1.

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, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. *arXiv preprint*, cs.CL/2004.13922v1.

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/1906.08101v2.

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2017. Attention-over-attention neural networks for reading comprehension. In *Proceedings of the ACL*, pages 593–602, Vancouver, Canada.

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 *Proeedings of the NAACL-HLT*, pages 4171–4186, Minneapolis, MN.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2017. Gated-attention readers for text comprehension. In *Proceedings of the ACL*, pages 1832–1846, Vancouver, Canada.

Bhuwan Dhingra, Danish Pruthi, and Dheeraj Rajagopal. 2018. Simple and effective semi-supervised question answering. In *Proceedings of the NAACL-HLT*, pages 582–587.

Peng Ding and Xiaobing Zhou. 2018. Ynu deep at semeval-2018 task 12: A bilstm model with neural attention for argument reasoning comprehension. In *Proceedings of the SemEval*, pages 1120–1123, New Orleans, LA.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Proceedings of the NeurIPS*, pages 13063–13075.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new Q&A dataset augmented with context from a search engine. *arXiv preprint*, cs.CL/1704.05179v3.

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

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the EMNLP*, pages 489–500.

Alexander Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. Template-based question generation from retrieved sentences for improved unsupervised question answering. In *Proceedings of the ACL*, pages 4508–4513.

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*, pages 1774–1784, Sofia, Bulgaria.

Maxwell Forbes and Yejin Choi. 2017. Verb physics: Relative physical knowledge of actions and objects. In *Proceedings of the ACL*, pages 266–276, Vancouver, Canada.

David Golub, Po-Sen Huang, Xiaodong He, and Li Deng. 2017. Two-stage synthesis networks for transfer learning in machine comprehension. *arXiv preprint*, cs.CL/1706.09789v3.

Ralph Grishman. 2019. Twenty-five years of information extraction. *Natural Language Engineering*, 25(6):677–692.

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.

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*, Caribe Hilton, Puerto Rico.

Lynette Hirschman, Marc Light, Eric Breck, and John D. Burger. 1999. Deep read: A reading comprehension system. In *Proceedings of the ACL*, pages 325–332, College Park, Maryland, USA. Association for Computational Linguistics.

Danqing Huang, Jing Liu, Chin-Yew Lin, and Jian Yin. 2018. Neural math word problem solver with reinforcement learning. In *Proceedings of the COLING*, pages 213–223, Santa Fe, NM.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019a. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the EMNLP-IJCNLP*, pages 2391–2401.

Zixian Huang, Yulin Shen, Xiao Li, Yuang Wei, Gong Cheng, Lin Zhou, Xinyu Dai, and Yuzhong Qu. 2019b. 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.

Stanley E Jones and Curtis D LeBaron. 2002. Research on the relationship between verbal and nonverbal communication: Emerging integrations. *Journal of communication*, 52(3):499–521.

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 Kočiskỳ, 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.

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.

Chia-Hsuan Lee, Shang-Ming Wang, Huan-Cheng Chang, and Hung-Yi Lee. 2018. ODSQA: Open-domain spoken question answering dataset. *arXiv preprint*, cs.CL/1808.02280v1.

Wendy Grace Lehnert. 1977. *The process of question answering*. Ph.D. thesis, Yale University.

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.

Patrick Lewis, Ludovic Denoyer, and Sebastian Riedel. 2019. Unsupervised question answering by cloze translation. In *Proceedings of the ACL*, pages 4896–4910.

Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. 2018a. Spoken SQuAD: A study of mitigating the impact of speech recognition errors on listening comprehension. *arXiv preprint*, cs.CL/1804.00320v1.

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. 2018b. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the ACL*, pages 138–143, Melbourne, Australia.

Yongbin Li and Xiaobing Zhou. 2018. Lyb3b at semeval-2018 task 11: Machine comprehension task using deep learning models. In *Proceedings of the SemEval*, pages 1073–1077, New Orleans, LA.

Zhongli Li, Wenhui Wang, Li Dong, Furu Wei, and Ke Xu. 2020. Harvesting and refining question-answer pairs for unsupervised QA. In *Proceedings of the ACL*, pages 6719–6728.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating Wikipedia by summarizing long sequences. In *Proceedings of the ICLR*, Vancouver, Canada.

Rui Liu, Junjie Hu, Wei Wei, Zi Yang, and Eric Nyberg. 2017. Structural embedding of syntactic trees for machine comprehension. In *Proceedings of the EMNLP*, pages 815–824.

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.

Kaixin Ma, Tomasz Jurczyk, and Jinho D Choi. 2018. Challenging reading comprehension on daily conversation: Passage completion on multiparty dialog. In *Proceedings of the NAACL-HLT*, pages 2039–2048, New Orleans, LA.

Michael McCandless, Erik Hatcher, and Otis Gospodnetic. 2010. *Lucene in Action, Second Edition: Covers Apache Lucene 3.0*. Manning Publications Co., Greenwich, CT.

Danielle S McNamara, Irwin B Levinstein, and Chutima Boonthum. 2004. iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments, & Computers*, 36(2):222–233.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the EMNLP*, pages 2381–2391.

George Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Sewon Min, Minjoon Seo, and Hannaneh Hajishirzi. 2017. Question answering through transfer learning from large fine-grained supervision data. In *Proceedings of the ACL*, pages 510–517, Vancouver, Canada.

Paramita Mirza and Raffaella Bernardi. 2013. Ccg categories for distributional semantic models. In *Proceedings of the RANLP*, pages 467–474, Hissar, Bulgaria.

Kouider Mokhtari and Carla A Reichard. 2002. Assessing students' metacognitive awareness of reading strategies. *Journal of educational psychology*, 94(2):249.

Kouider Mokhtari and Ravi Sheorey. 2002. Measuring esl students' awareness of reading strategies. *Journal of developmental education*, 25(3):2–11.

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.

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.

I Nation. 2006. How large a vocabulary is needed for reading and listening? *Canadian modern language review*, 63:59–82.

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.

Jianmo Ni, Chenguang Zhu, Weizhu Chen, and Julian McAuley. 2018. Learning to attend on essential terms: An enhanced retriever-reader model for open-domain question answering. In *Proceedings of the NAACL-HLT*, Minneapolis, MN.

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.

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2016. Who did What: A large-scale person-centered cloze dataset. In *Proceedings of the EMNLP*, pages 2230–2235, 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.

Soham Parikh, Ananya Sai, Preksha Nema, and Mitesh M Khapra. 2018. Eliminet: A model for eliminating options for reading comprehension with multiple choice questions. In *Proceedings of the IJCAI-ECAI*, pages 4272–4278, Stockholm, Sweden.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12:2825–2830.

Anselmo Penas, Yusuke Miyao, Alvaro Rodrigo, Eduard H Hovy, and Noriko Kando. 2014. Overview of CLEF QA Entrance Exams Task 2014. In *Proceedings of the CLEF*, pages 1194–1200, Sheffield, UK.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the NAACL-HLT*, pages 2227–2237, New Orleans, LA.

Rajkumar Pujari and Dan Goldwasser. 2019. Using natural language relations between answer choices for machine comprehension. In *Proceedings of the NAACL-HLT*, pages 4010–4015.

David D Qian and Mary Schedl. 2004. Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing*, 21(1):28–52.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. In *Preprint*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of th 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 JC 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.

Ellen Riloff and Michael Thelen. 2000. A rule-based question answering system for reading comprehension tests. In *ANLP-NAACL 2000 Workshop: Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*.

Mrinmaya Sachan, Kumar Dubey, Eric Xing, and Matthew Richardson. 2015. Learning answer-entailing structures for machine comprehension. In *Pro-*

ceedings of the ACL-IJCNLP, pages 239–249, Beijing, China. Association for Computational Linguistics.

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.

Stuart C Shapiro. 1992. *Encyclopedia of Artificial Intelligence, Second Edition*. New York: John Wiley.

Hideyuki Shibuki, Kotaro Sakamoto, Yoshinobu Kano, Teruko Mitamura, Madoka Ishioroshi, Kelly Y Itakura, Di Wang, Tatsunori Mori, and Noriko Kando. 2014. Overview of the NTCIR-11 QA-Lab Task. In *NTCIR*.

Richard Sikes. 2007. Fuzzy matching in theory and practice. *Multilingual*, 18(6):39–43.

Ellery Smith, Nicola Greco, Matko Bošnjak, and Andreas Vlachos. 2015. A strong lexical matching method for the machine comprehension test. In *Proceedings of the EMNLP*, pages 1693–1698.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the AAAI*, pages 4444–4451.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *Proceedings of the EMNLP*, pages 4208–4219, Brussels, Belgium.

Saku Sugawara, Yusuke Kido, Hikaru Yokono, and Akiko Aizawa. 2017a. Evaluation metrics for machine reading comprehension: Prerequisite skills and readability. In *Proceedings of the ACL*, pages 806–817, Vancouver, Canada.

Saku Sugawara, Hikaru Yokono, and Akiko Aizawa. 2017b. Prerequisite skills for reading comprehension: Multi-perspective analysis of mctest datasets and systems. In *Proceedings of the AAAI*, pages 3089–3096, San Francisco, CA.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, and Claire Cardie. 2020a. Improving machine reading comprehension with contextualized commonsense knowledge. *arXiv preprint*, cs.CL/2009.05831v2.

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

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2019b. Improving machine reading comprehension with general reading strategies. In *Proceedings of the NAACL-HLT*, pages 2633–2643.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2019c. Probing prior knowledge needed in challenging chinese machine reading comprehension. *arXiv preprint*, cs.CL/1904.09679v2.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2020b. Investigating prior knowledge for challenging chinese machine reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:141–155.

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

Chenhao Tan and Lillian Lee. 2015. All who wander: On the prevalence and characteristics of multi-community engagement. In *Proceedings of the WWW*, pages 1056–1066, Florence, Italy.

Niket Tandon, Gerard De Melo, Abir De, and Gerhard Weikum. 2015. Knowlywood: Mining activity knowledge from hollywood narratives. In *Proceedings of the CIKM*, pages 223–232.

Niket Tandon, Gerard De Melo, Fabian Suchanek, and Gerhard Weikum. 2014. Webchild: Harvesting and organizing commonsense knowledge from the web. In *Proceedings of the WSDM*, pages 523–532.

Min Tang, Jiaran Cai, and Hankz Hankui Zhuo. 2019. Multi-matching network for multiple choice reading comprehension. In *Proceddings of the AAAI*, Honolulu, HI.

Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2018. Multi-range reasoning for machine comprehension. *arXiv preprint*, cs.CL/1803.09074v1.

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.

Bo-Hsiang Tseng, Sheng-Syun Shen, Hung-Yi Lee, and Lin-Shan Lee. 2016. Towards machine comprehension of spoken content: Initial toefl listening comprehension test by machine. In *Proceedings of the Interspeech*, San Francisco, CA.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. We usually don't

like going to the dentist: Using common sense to detect irony on twitter. *Computational Linguistics*, 44(4):793–832.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the NIPS*, pages 5998–6008, Long Beach, CA.

Chao Wang and Hui Jiang. 2019. Explicit utilization of general knowledge in machine reading comprehension. In *Proceedings of the ACL*, pages 2263–2272.

Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2015. Machine comprehension with syntax, frames, and semantics. In *Proceedings of the ACL*, pages 700–706.

Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, Dong Yu, David McAllester, and Dan Roth. 2019. Evidence sentence extraction for machine reading comprehension. In *Proceedings of the CoNLL*, pages 696–707.

Lei Wang, Yan Wang, Deng Cai, Dongxiang Zhang, and Xiaojiang Liu. 2018a. Translating a math word problem to a expression tree. In *Proceedings of the EMNLP*, pages 1064–1069, Brussels, Belgium.

Liang Wang, Sujian Li, Wei Zhao, Kewei Shen, Meng Sun, Ruoyu Jia, and Jingming Liu. 2018b. Multi-perspective context aggregation for semi-supervised cloze-style reading comprehension. In *Proceedings of the COLING*, pages 857–867.

Liang Wang, Meng Sun, Wei Zhao, Kewei Shen, and Jingming Liu. 2018c. Yuanfudao at SemEval-2018 task 11: Three-way attention and relational knowledge for commonsense machine comprehension. In *Proceedings of the SemEval*, pages 758–762.

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

Yong Wang, Longyue Wang, Shuming Shi, Victor OK Li, and Zhaopeng Tu. 2020. Go from the general to the particular: Multi-domain translation with domain transformation networks. In *Proceedings of the AAAI*, pages 9233–9241.

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

Georg Wiese, Dirk Weissenborn, and Mariana Neves. 2017. Neural domain adaptation for biomedical question answering. In *Proceedings of the CoNLL*, pages 281–289, Vancouver, Canada.

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*, pages 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 2344–2356, Brussels, Belgium.

Frank F. Xu, Bill Yuchen Lin, and Kenny Zhu. 2018a. Automatic extraction of commonsense LocatedNear knowledge. In *Proceedings of the ACL*, pages 96–101.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi,

Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A chinese language understanding evaluation benchmark. In *Proceedings of the COLING*, pages 4762–4772.

Yichong Xu, Jingjing Liu, Jianfeng Gao, Yelong Shen, and Xiaodong Liu. 2018b. Dynamic fusion networks for machine reading comprehension. *arXiv preprint*, cs.CL/1711.04964v2.

An Yang, Quan Wang, Jing Liu, Kai Liu, Yajuan Lyu, Hua Wu, Qiaoqiao She, and Sujian Li. 2019. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. In *Proceedings of the ACL*, pages 2346–2357.

Zhilin Yang, Junjie Hu, Ruslan Salakhutdinov, and William Cohen. 2017. Semi-supervised QA with generative domain-adaptive nets. In *Proceedings of the ACL*, pages 1040–1050, Vancouver, Canada.

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.

Zhi-Xiu Ye, Qian Chen, Wen Wang, and Zhen-Hua Ling. 2019. Align, mask and select: A simple method for incorporating commonsense knowledge into language representation models. *arXiv preprint*, cs.CL/1908.06725v5.

Wen-tau Yih, Ming-Wei Chang, Christopher Meek, and Andrzej Pastusiak. 2013.

Question answering using enhanced lexical semantic models. In *Proceedings of the ACL*, pages 1744–1753, Sofia, Bulgaria.

Zhao You, Dan Su, and Dong Yu. 2019. Teach an all-rounder with experts in different domains. In *Proceedings of the ICASSP*, pages 6425–6429.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. QANet: Combining local convolution with global self-attention for reading comprehension. In *Proceedings of the ICLR*.

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020. Dialogue-based relation extraction. In *Proceedings of the ACL*, pages 4927–4940.

Xingdi Yuan, Tong Wang, Caglar Gulcehre, Alessandro Sordoni, Philip Bachman, Saizheng Zhang, Sandeep Subramanian, and Adam Trischler. 2017. Machine comprehension by text-to-text neural question generation. In *Proceedings of the RepL4NLP*, pages 15–25.

Hongming Zhang, Daniel Khashabi, Yangqiu Song, and Dan Roth. 2020. Transomcs: From linguistic graphs to commonsense knowledge. In *Proceedings of the IJCAI*.

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.

Shiyue Zhang and Mohit Bansal. 2019. Addressing semantic drift in question generation for semi-supervised question answering. In *Proceedings of the EMNLP-IJCNLP*, pages 2495–2509.

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.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal common sense acquisition with minimal supervision. In *Proceedings of the ACL*, pages 7579–7589.

Haichao Zhu, Li Dong, Furu Wei, Wenhui Wang, Bing Qin, and Ting Liu. 2019. Learning to ask unanswerable questions for machine reading comprehension. In *Proceedings of the ACL*, pages 4238–4248.

Haichao Zhu, Furu Wei, Bing Qin, and Ting Liu. 2018. Hierarchical attention flow for multiple-choice reading comprehension. In *Proceedings of the AAAI*, pages 6077–6084, New Orleans, LA.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE ICCV*, pages 19–27, Santiago, Chile.