A New Multi-choice Reading Comprehension Dataset for Curriculum Learning

Yichan Liang LYCH8@MAIL2.SYSU.EDU.CN
Jianheng Li LIJHENG3@MAIL2.SYSU.EDU.CN
Jian Yin* ISSJYIN@MAIL.SYSU.EDU.CN
School of Data and Computer Science, Sun Yat-sen University, China.
Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou 510006, P.R.China

Abstract

The past few years have witnessed the rapid development of machine reading comprehension (MRC), especially the challenging sub-task, multiple-choice reading comprehension (MCRC). And the release of large scale datasets promotes the research in this field. Yet previous methods have already achieved high accuracy of the MCRC datasets, *e.g.* RACE. It's necessary to propose a more difficult dataset which needs more reasoning and inference for evaluating the understanding capability of new methods. To respond to such demand, we present RACE-C, a new multi-choice reading comprehension dataset collected from college English examinations in China. And further we integrate it with RACE-M and RACE-H, collected by Lai et al. (2017) from middle and high school exams respectively, to extend RACE to be RACE++. Based on RACE++, we propose a three-stage curriculum learning framework, which is able to use the best of the characteristic that the difficulty level within these three sub-datasets is in ascending order. Statistics show the higher difficulty level of our collected dataset, RACE-C, compared to RACE's two sub-datasets, *i.e.*, RACE-M and RACE-H. And experimental results demonstrate that our proposed three-stage curriculum learning approach improves the performance of the machine reading comprehension model to an extent.

Keywords: Machine Reading Comprehension, Curriculum Learning

1. Introduction

In the age of information explosion, AI system in the field of natural language processing and understanding help people grab useful information from massive data efficiently and accurately, which requires the machine to read and comprehend, resulting on the task of Machine Reading Comprehension (MRC)(Hermann et al. (2015); Rajpurkar et al. (2016); Choi et al. (2018); Reddy et al. (2019)). The goal of MRC tasks is to have machines read a text passage and then make answers (a text span of the passage, a generated answer or a choice from several candidates) given any question about the passage. An AI agent which can display such capabilities would be useful in a wide variety of commercial applications such as answering general knowledge queries from Wikipedia documents, answering questions from financial reports of a company, troubleshooting using product manuals, *etc.*

* Corresponding author

Thanks to the rapid release of various large-scale datasets, machine reading comprehension (MRC) has been studied extensively in the literature. Taking the format of the datasets into account, MRC can be divided into three categories, namely cloze-style MRC(such as CNN/DailyMail(Hermann et al. (2015)), Children's Book Test (CBT)(Hill et al. (2015))), span-extraction MRC (such as SQuAD(Rajpurkar et al. (2016))), and multiple-choice MRC (such as MCTest(Richardson et al. (2013)), RACE(Lai et al. (2017))).

In this paper, we mainly focus on solving Multiple-Choice Reading Comprehension (MCRC), aiming at selecting the correct answer from a set of candidates given a question and a passage. At the beginning of the reading comprehension study, this type of reading comprehension task was not that popular because there is no large-scale dataset available and thus we can not apply neural network approaches to solve them. To bring more challenges to reading comprehension task and mitigate the absence of large-scale multi-choice reading comprehension dataset, Lai et al. (2017) proposed a novelty dataset called RACE. The RACE dataset is collected from the English examinations for Chinese middle and high school students, consisting of 27,933 passages and 97,687 questions. The large size of this dataset makes it possible to train and evaluate complex neural network based models and measure the scientific progress on MCRC. In addition, the latest breakthrough in Natural Language Processing (NLP): XLnet(Yang et al. (2019)) achieves the new state-of-the-art results on RACE dataset to 81.75%, outperforming Amazon Mechanical Turker for both RACE-M and RACE-H (two subsets of RACE) simultaneously for the first time. To be honest, we suspect the quality of RACE, or specifically, we doubt the rationality of the process of data cleaning. We found that (1) questions that contain keywords "underline" were not be removed although Lai et al. (2017) declared they have remove questions containing keywords "underlined" to avoid the irreproducibility of the effect of the underlines; (2) RACE is duplicated to a certain extent, even though the authors claimed that they have removed all the duplicated articles. As a result, we think that the performance of this dataset nearly reach its best level and the difficulty of this dataset is limit to Chinese middle and high school students' examination whose semantic information and grammatical structure have already been captured by several latest models(such as XLNet(Yang et al. (2019)), DCMN(Zhang et al. (2019)), OCN(Ran et al. (2019)), BERT(Devlin et al. (2018)), Reading Strategies Model(Sun et al. (2019b)), GPT(Radford et al.)). A more challenging dataset needs to be released to test the capacity of these recent models. Thus we propose a new dataset, RACE-C, extracted from college examinations in China. Compared with RACE, RACE-C is more challenging and needs more inference to answer the given questions, since the correct answer for most part of questions may not appear in the original passage directly and needs understanding of both natural language and world knowledge. In addition, we combine RACE and RACE-C into RACE++, a more integral dataset. And we hope that this new dataset can serve as a valuable resource for research and evaluation on machine reading comprehension (MRC). The dataset will be available at https://github.com/mrcdata/race-c/.

As we all know, external knowledge plays a critical role in human reading and understanding since authors assume readers have a certain amount of background knowledge gained from sources outside the text (Salmerón et al. (2006); Zhang and Seepho (2013)), so does RACE, which was designed carefully for evaluating Chinese middle and high school students (range between 12 to 18 years old)' ability in understanding and reasoning. Def-

initely, these students have their own knowledge systems from any other source ever since they have had their cognitive sense as young children. As a result, to replete with the knowledge gaps between humans and machines, very recent studies (Radford et al.; Devlin et al. (2018); Yang et al. (2019)) leverage rich world knowledge by pre-training deep neural models such as LSTMs and Transformers (Vaswani et al. (2017); Liu et al. (2018)) using language model objectives over large-scale corpora (*e.g.*, BookCorpus(Zhu et al. (2016)) and Wikipedia articles). We have seen significant improvements obtained on a wide range of natural language processing (NLP) tasks including MRC by fine-tuning these pre-trained general-purpose models on a downstream task. However, similar to the process of knowledge accumulation for human readers, it is relatively time-consuming and resource-extensive 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. (2016); Radford et al.); Devlin et al. (2018) 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 (about one year to train on eight P100 GPUs); Yang et al. (2019) train XLNet-Large on 512 TPU v3 chips for 500K steps with an Adam optimizer, linear learning rate decay and a batch size of 2048, which takes about 2.5 days. Without such strong devices support, we merely aim to fine-tune these excellent pre-trained model on our MCRC downstream task for the sake of introduction of external knowledge.

Inspired by curriculum learning(Bengio et al. (2009)) and self-paced learning(Kumar et al. (2010)), we proposed a new framework to train MCRC model. As Bengio et al. (2009) said, humans and animals learn much better when the examples are not randomly presented but organized in a meaningful order which illustrates gradually more concepts, and gradually more complex ones. Due to this human cognition, we assume that if model train middle and high examination step by step, the performance would be better. Thus we adapt these approaches of curriculum learning to our college new dataset, and as we expected, the performance of curriculum learning surpasses the result of mixed training by 2.6%.

In summary, our contributions consist of the following two parts:

1. We present RACE-C, a much more challenging MCRC dataset collected from the college English examinations, with which we extend RACE to be RACE++.

2. In consideration of the ascending difficulty of the subsets in RACE++, we propose a three-stage curriculum learning framework and demonstrate the performance of our method.

2. Related Work

2.1. Machine Reading Comprehension

Recently the progress in MRC is remarkable due to the introduction of large-scale datasets. We can divide the MRC tasks into three categories according to the answering formats of the datasets:

- **Cloze-style MRC** (Hermann et al. (2015); Hill et al. (2015); Onishi et al. (2016)) is to predict a missing word/entity in the question according to a passage. CNN/Daily

Mail(Hermann et al. (2015)) consists of 93k articles from the CNN as well as 220k articles from the Daily Mail websites and their corresponding bullet points as well as summarizing aspects of the information contained in the article. Notice that these summary points are abstractive and do not simply copy sentences from the articles. Hermann et al. (2015) construct a corpus of *(passage, query, answer)* triples by turning these bullet points into cloze(Taylor (1953)) style questions by replacing one entity at a time with a placeholder. Specifically, the passage is the news article, the query is its corresponding summary, of whom one entity is replaced with a placeholder, and the answer is the extracted entity. So the task is to cloze a question (summary) given a passage and the question (summary). So as Children's Book Test (CBT) (Hill et al. (2015)), where each passage consists of 20 contiguous sentences extracted from children's books and the 21st sentence is used to make the question. Who Did What (WDW) (Onishi et al. (2016)) is yet another cloze-style dataset constructed from the LDC English Gigaword newswire corpus. Onishi et al. (2016) generate passages and questions by picking two news articles describing the same event, using one as the passage and the other as the question.

- **Span-extraction MRC** (Rajpurkar et al. (2016); Nguyen et al. (2017)) is to predict the starting and ending indices in the passage, of which the answer is a text span. Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al. (2016)) consists of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, or the question might be unanswerable. MS MARCO(Nguyen et al. (2017)) comprises of anonymized questions sampled from Bing's search query logs, along with a human generated answer from passages extracted from web documents retrieved by Bing search engine. The answer to a certain question may not be unique and could be multiple spans, or no answer at all.

- **Multi-choice MRC** (Richardson et al. (2013); Lai et al. (2017); Sun et al. (2019a); Clark et al. (2018)) is to select the correct answer from a set of candidate answers given a question and an article. MCTest(Richardson et al. (2013)) is a high quality reading comprehension dataset for 7 years old children, containing 500 crowdsourcing stories and 2,000 questions, where each question is followed by four candidate answers and only one of them is correct. But the size of this dataset is too small to efficiently train advanced machine comprehension models (*i.e.* neural network models). Thus the majority of previous works on MCTest are feature-engineering models (Richardson et al. (2013); Narasimhan and Barzilay (2015); Smith et al. (2015); Wang et al. (2015)), which strongly rely on lexical, syntactic and frame semantic features extracted by various NLP tools. RACE(Lai et al. (2017)) is in the same format as MCTest but of much larger size and higher difficulty, which consists of 27,933 passages and 97,687 questions collected from English exams designed for Chinese middle and high school students, age from 12 to 18. Lai et al. (2017) build a rule-based baseline(Richardson et al. (2013)) with sliding window algorithm and adapt the Stanford AR(Chen et al. (2016)) and the GA Reader(Dhingra et al. (2017a)) to RACE as strong neural baseline. DREAM(Sun et al. (2019a)) is a multiple-choice dialogue-based reading comprehension examination dataset consisting of 10,197 multiple-choice questions for 6,444

dialogues, collected from English-as-a-foreign-language examinations designed by human experts. AI2 Reasoning Challenge (ARC) (Clark et al. (2018)) is also collected from human tests but covers the field of grade-school (range from 3rd grade to 9th, *i.e.*, students typically of age 8 through 13 years) natural science. It contains 7787 science questions, all non-diagram, multiple choice (typically 4-way multiple choice), sorted into a challenge set of 2590 hard questions (those that both a retrieval and a co-occurrence method fail to answer correctly) and an easy set of 5197 questions.

In this paper, our work primarily focuses on multiple-choice examination datasets designed by educational experts (Lai et al. (2017); Clark et al. (2018); Sun et al. (2019a)) since questions from these datasets are generally clean, error-free and challenging. Since the current MCRC datasets is either too small in size (*e.g.* Richardson et al. (2013); Clark et al. (2018)) or nearly reach its best performance, we propose a new dataset collected from various college English examinations named RACE-C, inheriting the same data format as RACE(Lai et al. (2017)). Our dataset is more difficult and needs more inference, which can test the MCRC models' capability better.

As for the model, our work follows the general framework of discriminatively fine-tuning pre-trained language models on question answering tasks (Radford et al.; Devlin et al. (2018); Sun et al. (2019b); Yang et al. (2019)). As a matter of fact, researchers develop a variety of methods with attention mechanisms (Chen and Choi (2016); Dhingra et al. (2017b); 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)). These methods seldom take the rich external knowledge (other than pre-trained word embeddings) into considerations. Instead, we investigate different approaches based on an existing pre-trained Transformer(Devlin et al. (2018)) (Section 4.1), which leverages rich linguistic knowledge from external corpora and achieves state-of-the-art performance on a wide range of NLP tasks including machine reading comprehension.

2.2. Curriculum Learning

Curriculum learning is a learning framework proposed by Bengio et al. (2009), in which a model is learned by gradually training from easy to complex samples so as to increase the entropy of training samples(Bengio et al. (2009)). Specifically, we define a function $f(x)$ for each sample x in the dataset. The value of $f(x)$ represents the difficulty of the sample x. As a result, we'd better train the model on dataset with smaller value of $f(x)$ first, and then train on the bigger ones. In curriculum learning, $f(x)$ is often predefined by humans' prior knowledge. Moreover, Spitkovsky et al. (2010) assume that short sentence is easier to understand than long sentence, which signifies that longer sentences mean higher level of difficulty.

In addition, there is a similar framework inherited from curriculum learning, called self-paced learning, whose ranking function $f(x)$ is defined by the output of the model rather than derived by predetermined heuristics. And it also choose the simple samples each time, while active learning, another learning framework, do the opposite. In active learning, a dataset is partitioned into important samples and unimportant ones. Important samples exert considerable influence on the training of a model. Actually, important samples tend to be the hard samples. We are eager to find out a few important samples among a large

Sample from RACE-C(our dataset)	Sample from RACE
Passage: Students of United States history, seeking to identify the circumstances that encouraged the emergence of feminist movements, have thoroughly investigated the mid-nineteenth-century American economic and social conditions that affected the status of women. ... ... The envisioned result of both currents of thought, however, was that women would enter public life in the new age and that sexual equality would reward men as well as women with an improved way of life.	**Passage**: There are few families in the United States that do not have either a radio or television set. Both of them have become a necessary part of our daily life, keeping us filled with the news of the day, teaching us in many fields of interest, and making us happy with singing, dancing and acting. ... ... Now a family in Chicago can watch on TV a motor-car race in Italy, a table tennis competition in Beijing or a volleyball match in Japan as these events are actually happening!
Questions: 1. It can be inferred that the author considers those historians who describe early feminists in the United States as "solitary" to be _ . **A. insufficiently familiar with the international origins of nineteenth-century American feminist thought** B. overly concerned with the regional diversity of feminist ideas in the period before 1848 C. not focused narrowly enough in their geo-graphical scope D. insufficiently aware of the ideological consequences of the Seneca Falls conference	**Questions**: 1.The passage tells us that _ in the U.S.A. have no radio or television set. **A. few families** B. all the families C. many families D. a few families 2. Who do you think the writer of the passage is? A. An Italian. B. A Japanese. **C. An American.** D. A Chinese.

Table 1: Sample reading comprehension problems from RACE-C(our dataset) and RACE.

dataset and train them to get a mature model which achieves a similar accuracy as the one trained with all samples of the full dataset. Similar to self-paced learning, self-training tends to select the easier samples. However, self-training is a semi-supervised learning method, and it need to predict the difficulty of samples without annotations.

3. Our Dataset

In this section, we describe how we construct our dataset RACE-C (Section 3.1) and provide a detailed analysis of this dataset (Section 3.2). In addition, we present a vertical comparison with RACE(Lai et al. (2017)), *i.e.*, RACE-M and RACE-H, the precursor of RACE-C.

3.1. Collection Methodology

We collect our college reading comprehension dataset from a variety of English examinations (including practice examinations) such as College English Test, Test for English Majors, English for Professional Titles and Public English Test[1], which are designed by English instructors to assess the reading comprehension level of Chinese English learners in college (typically aged 18-24). Following the naming rules of RACE, where RACE-M denotes the middle school examinations and RACE-H denotes high school examinations, we call our college examinations RACE-C. The topic coverage of RACE-C is broad and the content

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

Dataset	RACE-M				RACE-H				RACE-C(Ours)			
Subset	Train	Dev	Test	All	Train	Dev	Test	All	Train	Dev	Test	All
# passages	6409	368	362	7139	18728	1021	1045	20794	2437	136	135	4275
# questions	25421	1436	1436	28293	62445	3451	3498	69394	12702	712	708	14122

Table 2: The separation of the training, development, and test sets in RACE-M, RACE-H and RACE-C.

is all-encompassing, such as animals, plants, biographies, history, culture, environment, resources, transportation, medicine, economy, and information, *etc.* All the problems in RACE-C are freely accessible online for public usage. Each problem consists of a passage and a series of multiple-choice questions. The data before cleaning contains 4,451 passages and 22,692 questions.

We conduct the following filtering steps to clean the raw data. Firstly, we remove all the problems whose number of questions mismatch the number of options list or answers list, *e.g.*, problem with 5 questions but 4 answers or 4 options lists will be removed completely. Problems only with the same number of questions, options lists and answers will be retained. Secondly, we remove all problems and questions that do not have the same format as our problem setting, *e.g.*, a question would be removed if the number of its options is not four. Thirdly, we filter all articles and questions that are not self-contained based on the text information, *i.e.* we remove the articles and questions containing images or tables. We also remove all questions containing keywords "underline" or "underlined", since it is difficult to reproduce the effect of underlines. But we didn't remove questions containing keywords "paragraph" for we expect models could capture the paragraph segment information from the article. Fourthly, we remove all duplicated articles. Finally, we get the cleaned dataset RACE-C with 4,275 articles and 14,122 questions. A sample from our dataset is presented on the left of Table 1.

3.2. Data Analysis

We summarize data split in Table 2 and the statistics of RACE-C in Table 3, summing up the data partition and statistics of RACE (RACE-M and RACE-H), respectively. Following RACE, we split 5% data as the development set and 5% as the test set. As shown in Table 3, the average number of sentences or words of articles in RACE-M, RACE-H, RACE-C is in increasing order, so as the average number of words of questions and options, proving our recognition, or the fact, that the difficulty of English exams in middle school, high school, and college keep a gradual increasing trend. What's more, the total number of words in all the dataset, RACE-C, is 1,727,117 while the total number of words in RACE-M is 2,497,893. But the vocabulary size of RACE-C is 58,812 while the metric of RACE-M is 38,564. In other words, the size of vocabulary of RACE-C is 1.5 times of that of RACE-M, while the size of the complete RACE-C is seven tenths of that of RACE-M. By the way, the scale of RACE-H is too large to be compared with RACE-C reasonably, so we use Equation 1 to evaluate the non-repetition rate of RACE-C and RACE-H as well as RACE-M.

$$rate = \frac{S_C - S_X}{S_C} \tag{1}$$

Dataset	RACE-M	RACE-H	RACE-C(Ours)
min/avg/max sentences of article	1/16.6/63	1/18.0/111	2/18.3/68
min/avg/max words of article	2/232.1/626	3/354.1/1391	53/416.3/1359
min/avg/max words of question	1/10.0/75	1/11.4/76	2/13.2/65
min/avg/max words of options	1/4.9/38	1/6.8/115	1/7.3/40
min/avg/max questions per article	1/4.0/7	0/3.3/6	1/5.2/13
total # of words	2,497,893	10,041,248	1,727,117
vocabulary size	38,564	143,639	58,812

Table 3: The overall statistics of RACE-M, RACE-H and RACE-C.

S_C denotes the vocabulary size of RACE-C, while S_X can be specified to S_M or S_H to represent the vocabulary size of RACE-M or RACE-H, respectively. By applying this operating rule, we get the non-repetition rate of RACE-C with RACE-H, 38.5%, which illustrates that there are still a bit part of words that don't appear in the vocabulary of RACE-H, even though the scale of vocabulary in RACE-H is much more larger (about 2.4 times the vocabulary size of RACE-C). In addition, the non-repetition rate of RACE-C with RACE-M is 70.1%, demonstrating that most of the words in RACE-C (college dataset) are of higher level and higher difficulty, and do not appear in RACE-M (middle school dataset). These statistics strongly proof that our dataset, RACE-C, is naturally more difficult than RACE (including RACE-M and RACE-H) and needs more reference ability. An intuitive example of RACE is shown on the right of Table 1.

By the way, we found that the data in RACE(Lai et al. (2017)) is not that clean as the author declared in the paper. From the statistics (Table 3) we found that there are some articles having no questions, shown as the zero value of the min questions per article. And many articles have only one sentence. Thus we open some files with one article-sentence, in which we found that some of the articles is of the type of information matching of Chinese students English exams, which is not appropriate for the current task we research now. Besides, we found that questions containing keywords "underline" were not be removed although Lai et al. (2017) declared that they have remove questions containing keywords "underlined" to avoid the irreproducibility of the effect of the underlines. It will somehow effect the data quality. In addition, the samples in RACE are duplicated to a certain extent. In a word, the dataset is not that clean as Lai et al. (2017) declared. So we re-clean the data in RACE and sums our cleaned RCAE-C up to integrate into RACE++, a large-scale reading comprehension dataset covering middle school, high school and college's English examinations in China.

We conduct human annotations of questions types to get a comprehensive picture about the reasoning difficulty requirement of RACE-C. Following Lai et al. (2017), we classify the questions into five classes below:

- Word matching: The question is a text span in the article, and the answer is transparent.

- Paraphrasing: The question is paraphrased by exactly one sentence of the article, and the answer can be extracted within the sentence.

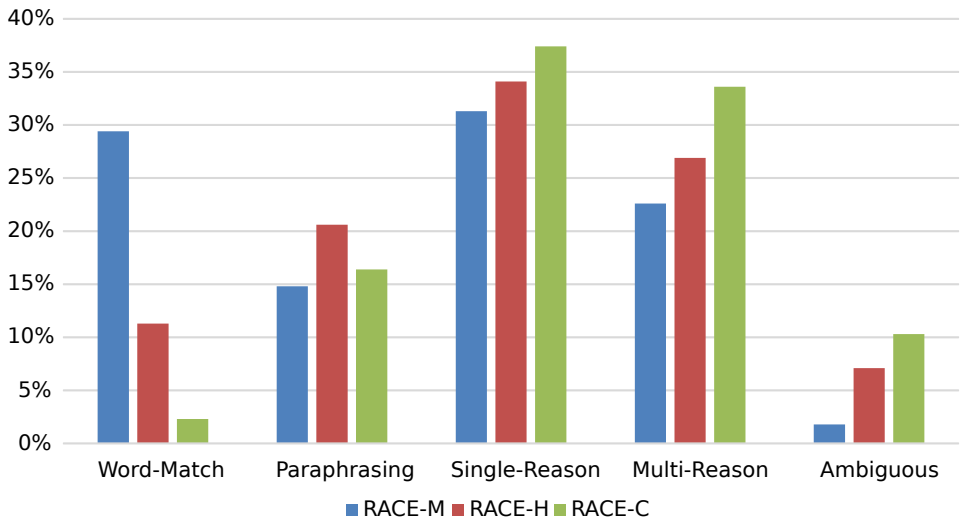

Figure 1: Visualization of statistic information about Reasoning type in different datasets.

- Single-sentence reasoning: The answer could be inferred from a single sentence of the article by recognizing incomplete information or conceptual overlap.
- Multi-sentence reasoning: The answer must be inferred from synthesizing information distributed across multiple sentence.
- Insufficient/Ambiguous: The question has no answer or the answer is not unique based on the given passage.

The difficulty of the five reasoning types is in ascending order. We sample 200 questions from RACE-C to obtain the proportion of each question types, and the statistics, compared with RACE (RACE-M, RACE-H), is summarized in Figure 1. 71.0% questions of RACE-C are of reasoning type (single-sentence reasoning and multi-sentence reasoning), while the proportion of RACE-M and RACE-H is 53.9% and 61.0% respectively. Also notice that the proportion of word matching questions on RACE-C is only 2.3%, the lowest among the three, while the proportion of RACE-H (11.3%) is lower than that of RACE-M (29.4%). As we all know, the more reasoning the questions need, the higher level of difficulty they belong. Thus we can get an conclusion that RACE-C are the most complex since it has the highest proportion of reasoning questions and lowest proportion of word matching questions, followed by RACE-H and RACE-M. An intuitive comparison of their samples is presented in Table 1.

4. Our Method

In this section, we first introduce the underlying question answering baseline we use (Section 4.1). Then we present our framework (Section 4.2) based on this baseline model.

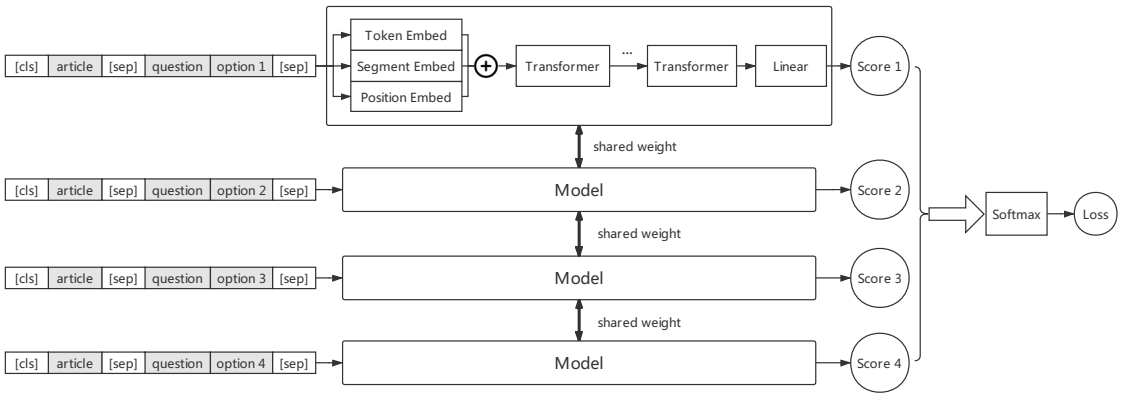

Figure 2: The overview of the basic model.

4.1. Basic Model

In this paper, we adopt BERT(Devlin et al. (2018)) as the pre-trained language model within our proposed framework. In the original BERT model, the input token sequence is either a single text sentence or a pair of text sentences. But in our MCRC dataset, there are three components of each example: passage, question and answers. So we input the passage as sentence A and the concatenation of question and option as sentence B. Since there are four options for each question (regarded as one example), we construct four input sequences for each question (example), each of the following format:

$$[\text{[CLS] Passage [SEP] Question } + \text{ Option [SEP]}]$$

We use the same special tokens, *i.e.* [CLS] and [SEP] as BERT, where the first one is always the special classification embedding, and the second is generally used to separate two sentences. The overview of our basic model is shown in Figure 2. The four token sequences contain four options of a question respectively. For a given token, its input representation is constructed by summing the corresponding token, segment and position embeddings. Each token sequence is required to go through a multi-layer bidirectional transformer encoder. Following the same fine-tuning procedure as Devlin et al. (2018), we use the final hidden vector corresponding to first input token ([CLS]) as the aggregation representation of the whole sequence. And then a linear layer is applied on the aggregation representations of the four sequences to obtain the scores. Finally, a softmax function is adopted to compute the cross-entropy loss in the training phase or the probabilities in the testing phase.

4.2. Framework of Curriculum Learning

We use BERT(Devlin et al. (2018)) baseline to verify the effectiveness of curriculum learning on machine reading comprehension, for which we propose a three-stage curriculum leaning framework. This framework needs datasets to be in different levels of comprehensive difficulty, and that's why we supplement our collected harder RACE-C to RACE(Lai et al.

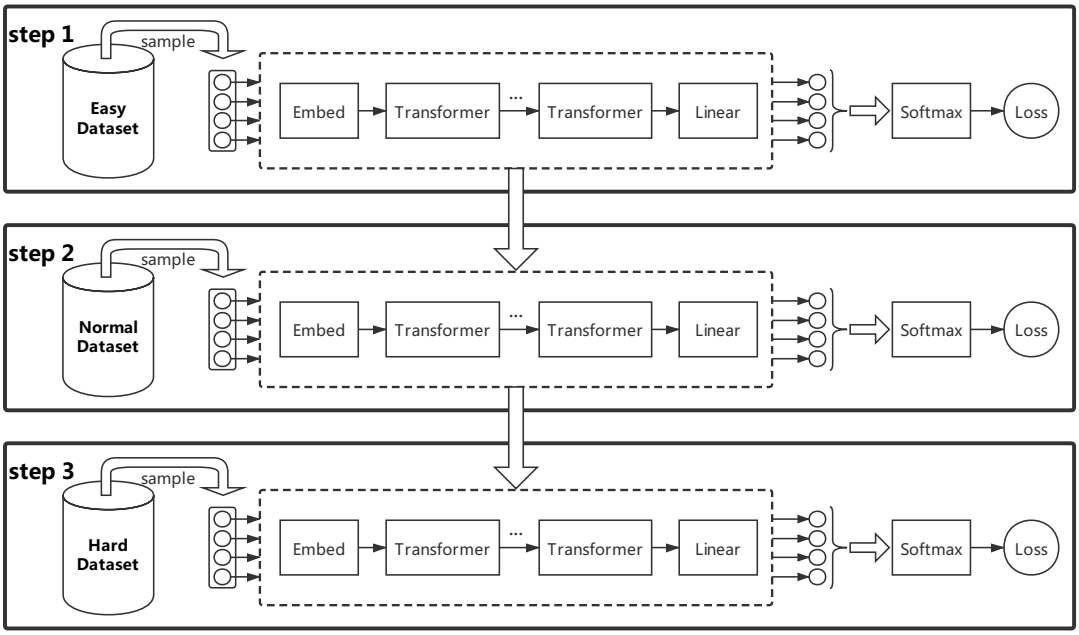

Figure 3: The overview of our proposed three-stage curriculum learning framework. For each step, we train the model on the corresponding dataset, and the trained parameters will be used as the initial parameters of the successive model.

(2017)) to be RACE++, including RACE-M, RACE-H and RACE-C with three levels of difficulty, corresponding to the three-stage curriculum leaning framework.

$$f(x) = \begin{cases} 0 & x \in \text{RACE-M} \\ 1 & x \in \text{RACE-H} \\ 2 & x \in \text{RACE-C} \end{cases} \tag{2}$$

The overall architecture of MCRC training framework based on curriculum leaning is depicted by Figure 3. The model used in this paper is BERT introduced in Section 4.1. We use the parameters pre-trained by BERT as our initial parameters. And we simply define the difficulty function of curriculum leaning as the partition of these three sub-datasets (see Equation 2). Specifically, our model will learn the easiest subset of RACE++, RCAE-M in step 1. And then it learns RACE-H, the middle level of difficulty among RACE-M and RACE-C, in step 2. Finally it learns our proposed dataset, RACE-C, the hardest subset of the three, in step 3.

As declared by Bengio et al. (2009), human learns much better when the examples are not mixedly presented but organized in a meaningful order which illustrates gradually more concepts, and gradually more complex ones. When model learns RACE-C, it is able to converge to a better result quickly since it has been equipped with the ability of doing

Dataset	Dev	Test
RACE-M	70.1	69.0
RACE-H	64.9	62.3
RACE-C	31.6	33.8

Table 4: Accuracy (%) of the fine-tuned BERT baseline on the three subsets of RACE++ dev and test set.

Method	Accuracy(RACE-C)
BERT baseline	33.8
after fine-tuned on RACE	45.5
three-stage learning method (ours)	**48.1**

Table 5: Accuracy (%) of BERT baseline method, the performance after fine-tuned on RACE, and our proposed three-stage curriculum learning method on RACE-C.

such reading comprehension task by learning the knowledge of RACE-M and RACE-H successively. On the contrary, if we learn the mixture of RACE-M and RACE-H first, then the result won't be so ideal (see Section 5.2).

Besides curriculum learning, the proposed dataset can be applied to other learning framework. For active learning, our dataset provide ternary-class (easy, normal, hard) coarse-grained label for a method based on active learning to evaluate if the model tends to select the hard samples. In contrast to active learning, self-paced learning can employ our ternary-class label to evaluate if it can select the easier samples. As for self-training, it tend to select samples of high confidence. Thus our provided ternary-class label can be used to analyze the relationship between the confidence and the difficulty of samples.

5. Experiments

5.1. Experimental Settings

We adopt the pre-trained uncased $\text{BERT}_{\text{LARGE}}$ released by Devlin et al. (2018) as our backbone model. $\text{BERT}_{\text{LARGE}}$ contains 24 hidden layers, and each layer contains 16 attention heads. The size of the word embedding vector is 1024. For less time consumption and less resource occupy, we use apex, maintained by NVIDIA to streamline mixed precision and distributed training in Pytorch, where the half-precision floating-point is the key operation to accelerate the calculation speed. We set the batch size to 8, learning rate to $1e-5$, and maximum sequence length to 320. We adopt Adam optimizer and fine-tune for 4 epochs on RACE-M, 2 epochs on RACE-H and 6 epochs on RACE-C, considering of the proportion of their respective number of questions. Thanks to the limitation of memory, we accumulate the gradient 8 times and update the parameters once for a batch, under the environment of a NVIDIA TITAN X GPU and Intel Xeon E5-2620 CPU.

5.2. Experimental Results

We show the accuracy of our implemented BERT baseline on the three sub-dataset (*i.e.* RACE-M, RACE-H, RACE-C) of RACE++ in Table 4 and our proposed approaches in Table 5.

As shown in Table 4, the fine-turned results of BERT on RACE-M, RACE-H and RACE-C are 69.0%, 62.3% and 33.8%, respectively, in the descending order, which indicates the ascending order of difficulty. On the other hand, the performance of our proposed three-stage curriculum learning method is presented in Table 5. We do a comparison experiment to verify the three-stage method's effectiveness. We first fine-tune BERT on the fusion of RACE (including RACE-M and RCAE-H) and then fine-tune on RACE-C, the final testing accuracy is lower than our proposed method by 2.6%, where we first fine-tune BERT on the easiest dataset RACE-M, then fine-turn on the normal dataset RACE-H, and finally fine-turn on our hardest dataset RACE-C. This experimental result effectively supports our assumption that neural network model can learn better by absorbing knowledge from simplicity to difficulty, and certainly demonstrate the performance of our proposed method.

5.3. Analysis and Discussion

As a matter of fact, not all examples of RACE-H are harder than that of RACE-M, so do RACE-C compared to RACE-H. Thus our Equation 2 merely gives a coarse estimation of the difficulty of each example according to its corresponding dataset. Maybe a more precise estimation is required to select samples and further improve the performance of our step-by-step curriculum learning framework. Self-paced learning seems a good choice despite of the complexity, which can ceaselessly find out easier samples from the untrained dataset during the training phase.

6. Conclusion

In this paper, we introduce a diversified, high-difficulty and high-quality dataset for machine reading comprehension that is carefully designed by English instructors to examine Chinese college students' ability on this task. Besides, we looked insight into our dataset, RACE-C, and made a deep contrast between RACE-C and RACE(Lai et al. (2017)) (including RACE-M and RACE-H). Further we combine these three sub-datasets into a complete dataset called RACE++. Inspired by the ascending difficulty of these three sub-datasets, we propose a three-stage curriculum learning framework, using the latest breakthrough neural network model(Devlin et al. (2018)) to train RACE++ step by step. Experimental results demonstrate that our three-stage curriculum learning approach outperforms the strategy of fused training RACE-M and RACE-H by 2.6%.

Acknowledgments

We would like to thank Jianxing Yu for suggestions on the idea, and Ruiying Zhou and Shiqi Wang's help on collecting part of the dataset. This work is supported by the Research Foundation of Science and Technology Plan Project in Guangdong Province and Guangzhou City (2015A030401057, 2016B030307002, 2014SY000013, 2017B030308007).

References

Yoshua Bengio, J Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM, 2009.

Danqi Chen, Jason Bolton, and Christopher D Manning. A thorough examination of the cnn/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, 2016.

Yu-Hsin Chen and Jinho D Choi. Character identification on multiparty conversation: Identifying mentions of characters in tv shows. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 90–100, 2016.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, 2018.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. Gated-attention readers for text comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1832–1846, 2017a.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. Gated-attention readers for text comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1832–1846, 2017b.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701, 2015.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*, 2015.

M Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, pages 1189–1197, 2010.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, 2017.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*, 2018.

Karthik Narasimhan and Regina Barzilay. Machine comprehension with discourse relations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1253–1262, 2015.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human generated machine reading comprehension dataset. In *International Conference on Learning Representations*, 2017.

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. Who did what: A large-scale person-centered cloze dataset. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2230–2235, 2016.

Soham Parikh, Ananya B Sai, Preksha Nema, and Mitesh M Khapra. Eliminet: a model for eliminating options for reading comprehension with multiple choice questions. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4272–4278. AAAI Press, 2018.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, 2016.

Qiu Ran, Peng Li, Weiwei Hu, and Jie Zhou. Option comparison network for multiple-choice reading comprehension. *arXiv preprint arXiv:1903.03033*, 2019.

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

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, 2013.

Ladislao Salmerón, Walter Kintsch, and José J Caãs. Reading strategies and prior knowledge in learning from hypertext. *Memory & Cognition*, 34(5):1157–1171, 2006.

Ellery Smith, Nicola Greco, Matko Bosnjak, and Andreas Vlachos. A strong lexical matching method for the machine comprehension test. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1693–1698, 2015.

Valentin I Spitkovsky, Hiyan Alshawi, and Daniel Jurafsky. From baby steps to leapfrog: How less is more in unsupervised dependency parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 751–759. Association for Computational Linguistics, 2010.

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

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. Improving machine reading comprehension with general reading strategies. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2633–2643, 2019b.

Min Tang, Jiaran Cai, and Hankz Hankui Zhuo. Multi-matching network for multiple choice reading comprehension. 2019.

Wilson L Taylor. "cloze procedure": A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433, 1953.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. Machine comprehension with syntax, frames, and semantics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 700–706, 2015.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.

Lian Zhang and Sirinthorn Seepho. Metacognitive strategy use and academic reading achievement: Insights from a chinese context. *Electronic Journal of Foreign Language Teaching*, 10(1), 2013.

Shuailiang Zhang, Hai Zhao, Yuwei Wu, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. Dual co-matching network for multi-choice reading comprehension. *arXiv preprint arXiv:1901.09381*, 2019.

Haichao Zhu, Furu Wei, Bing Qin, and Ting Liu. Hierarchical attention flow for multiple-choice reading comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. 2016.