A Thorough Examination of the RACE Machine Reading Comprehension Task

Wenlong Zhao
University of California, San Diego
Cognitive Science Department
Honors Program 2018-2019

Committee: Dr. Zhuowen Tu, Dr. Benjamin Bergen, Zhiyao Yan

Abstract

Machine reading comprehension (MRC) systems are typically evaluated by accuracy on the RACE MRC task. There is a significant gap between the state-of-the-art machine performance (74.1%) and the human ceiling performance (94.5%) on RACE, even though MRC systems have outperformed humans on a number of MRC tasks. In this paper, we aim at better understanding of the characteristics of RACE and the desiderata for human-level performance on RACE. On one hand, we categorize 300 questions randomly sampled from the RACE test set according to a set of prerequisite skills that we propose. On the other hand, we evaluate top MRC systems proposed for RACE on different categories of questions. The result suggests that existing systems have much room for improvement in terms of representing certain reasoning skills, such as relation identification and algebraic reasoning. Data categorization and code are available on github[1].

1 Introduction

Machine reading comprehension (MRC) is a long-term goal of natural language processing research. Burges (2013) suggests that we can operationally define MRC tasks as question answering (QA) tasks based on textual materials. QA is both a means of evaluating MRC systems and an application of MRC (Chen et al., 2018). A successful MRC system is expected to correctly answer questions about textual materials after processing them, as long as the questions are answerable by a majority of human readers of the materials who are proficient in the language.

There are four main types of MRC tasks, in the form of four types of text-based QA. In cloze style tasks (Hermann et al., 2015; Hill et al., 2015; Onishi et al., 2016), systems are expected to select tokens from sets of options or a vocabulary to complete sentences that contain placeholders. In span prediction tasks (Rajpurkar et al., 2016, 2018; Trischler et al., 2016; Joshi et al., 2017), systems are required to use token sequences in the original texts to answer questions about the texts. In free-form QA tasks (Nguyen et al., 2016; Kočiskỳ et al., 2018), systems may provide free-form answers, while the answering keys are often designed to be tokens or paraphrases of tokens from the original texts. In multiple choice tasks (Richardson et al., 2013; Lai et al., 2017), systems answer questions by choosing correct options from multiple candidate answers.

Neural network models proposed in recent years are capable of learning complex lexical matching through deep structures. These models perform well on questions that are solvable using only lexical cues. Such questions include most cloze style questions, span prediction questions, free-form questions, and some multiple choice questions. Machine systems that use neural network structures have outperformed humans on a number of tasks that consist of such questions[2].

On the other hand, it is realized that good performance on these tasks is not a sufficient indicator of human-level reading comprehension capacity (Clark et al., 2018). RACE is a multiple choice dataset claimed to involve much reasoning. Many questions in RACE cannot be answered using only lexical cues. Arguably in accordance is the fact that there exists a gap between the state-of-the-art model performance (74.1% test accuracy) and the human ceiling performance (94.5% test accuracy).

In this paper, we aim at understanding the

[1] https://github.com/wenlongzhao094/ExamineRACE

[2] Example MRC leaderboards where machine systems achieve better results than humans. Squad 1.1 and 2.0: https://rajpurkar.github.io/SQuAD-explorer; MS MARCO: http://www.msmarco.org/leaders.aspx.

desiderata for better MRC systems by analyzing the gap between the top machine performance and the human ceiling performance on RACE. Firstly, we analyze questions in RACE and propose a set of reasoning skills required for them. Then we categorize 300 questions randomly selected from the test set according to the proposed skills. Lastly we evaluate BERT$_{BASE}$, BERT$_{LARGE}$ (Devlin et al., 2018), and DCMN (based on BERT$_{BASE}$) (Zhang et al., 2019) on each category of questions. The models perform differently on different categories of questions, indicating that our data categorization is meaningful. The categorization may be effectively used to examine MRC models on representations of different reasoning types. While models achieve high accuracy on questions that are solvable through lexical matching, they perform much worse on questions that require certain reasoning skills, such as relation identification and algebraic reasoning. We conclude that building representations for these reasoning skills is the desiderata for human-level MRC.

2 Related Work

2.1 The RACE Dataset

The large-scale ReAding Comprehension dataset collected from English Examination (RACE) is a multiple choice style dataset (Lai et al., 2017). It consists of passages and multiple choice questions designed for middle and high school students who study English as a second language.

2.2 Analytical Approaches to MRC Datasets

Annotating questions with reasoning types is a common approach for understanding MRC datasets. The purpose of such annotation is usually to evaluate the overall difficulty of datasets. In the original RACE paper (Lai et al., 2017), the authors follow Chen et al. (2016) and Trischler et al. (2016) to stratify questions in RACE into five classes: word matching, paraphrasing, single-sentence reasoning, multi-sentence reasoning, and insufficient/ambiguous. These classes are, however, not indicative of the specific reasoning skills necessary for answering questions.

Lai et al. (2017) further list five reasoning categories observed in RACE: detail reasoning, whole-picture reasoning, passage summarizing, attitude analysis, and world knowledge. But the list is insufficient for the purpose of thorough examination of models on their representations for reasoning types. The definitions of ”detail” and ”world knowledge” are unclear. ”Whole picture reasoning” and ”summarizing” seem to overlap. The list is also not comprehensive enough to cover all reasoning skills required.

For the purpose of evaluating MRC systems based on reasoning skills, Sugawara et al. (2017) have proposed a list of reasoning skills for the MCTest dataset (Richardson et al., 2013):

- List/Enumeration
- Mathematical operations
- Coreference resolution
- Logical reasoning
- Analogy
- Spatiotemporal relations
- Causal relations
- Commonsense reasoning
- Schematic/Rhetorical clause relations
- Special sentence structure

Although the list involves some careful design, the authors do not sufficiently consider the nature of MRC. Consider the following example. Context: *John was annoyed because his sister ate his cake.* Question: *Why was John annoyed?* Answer: *Because his sister ate his cake.* Sugawara et al. (2017) states that this example involves causal relations. But indeed, the question can be easily solved with lexical matching. Also, the MCTest dataset is designed for 7-year-old children and involves less complex reasoning than RACE. The annotated MCTest dataset is insufficient to be used to examine model representations of reasoning skills.

3 Reasoning Skills in RACE

We propose eight categories of reasoning skills that are mutually exclusive and cover all questions in RACE. The proposal considers the nature of MRC and aims at operational classification of questions in RACE for the purpose of diagnosing the desiderata of better models. One example from each category, except for *Others*, is listed in Table 1.

Lexical Matching: The correct option appears in the passage or is paraphrasing lexical cues in the passage. There is no confusing options.

Inference: The correct option can be inferred from lexical cues in the passage. Representations for logical reasoning skills, such as, negation and deduction, are necessary for the machine to synthesize lexical cues in the text and obtain the correct option.

Passage: happiness is for everyone . ... **Exact Match:** happiness is for _ . A. those who have large and beautiful houses B. those who have cars C. those who have a lot of money **D. all people**
Passage: ... a part - time job can teach teenagers important skills like responsibility , independence , teamwork and leadership . ... **Inference:** the author thinks a part - time job can teach teenagers many skills except _ . A. responsibility B. teamwork C. leadership **D. organization**
Passage: ... these jobs are great for young people who want to be active and have fun while making money . a favorite job for many teens is babysitting , and they can start before 14 if the parents agree . after the children are sleeping and before the parents come home , babysitters have lots of freedom . as long as they stay in the house and make sure the kids are okay , babysitters can do their homework , enjoy a snack , watch tv , or talk on the phone with friends . ... **Relation Identification:** american teenagers like to babysit because they _ . A. can start before 14 B. can make more money C. like to play with children **D. have lots of freedom**
Passage: in britain , people often invite friends for a meal , a party or just coffee . people who know each other very well may visit each other ' s houses without an invitation , but if we invite new friends , usually an invitation is needed . ... these are usually just polite ways of ending a talk . ... **Summary:** which is the best title for the passage ? A. britain **B. invitation** C. a talk with friends D. a letter to friends
Coreference Resolution: in " at the end of it " , the word " it " means _ .
Passage: lucy ... four volleyballs ... mary ... five volleyballs ... **Algebraic Reasoning:** lucy and mary have _ volleyballs . A. four B. five C. eight **D. nine**
Passage: ... i started thinking about my dad coming home from work to find that i failed the test . . . " how could you have failed the test ? i am certain that nobody else in the whole class got as bad a grade as you did ! " ... **Commonsense Reasoning:** the writer ' s father was _ . A. proud of him B. tired of him C. **strict with him** D. pleased with him

Table 1: Examples from the test set of RACE that require different reasoning skills.

Relation Identification: Multiple options can be matched to lexical cues in the passage. The machine needs to track relations among entities and events in the passage to identify the most relevant and correct option. From human perspective, the relations could be conditional relations, causal relations, spatiotemporal relations, and so forth. From machines' perspective, all relations are different types of mapping between numerical representations that need to be learned.

Summary: Multiple options can be matched to lexical cues in the passage. The machine needs to identify the option that dominates the meaning of the passage from a certain aspect requested by the question.

Coreference Resolution: A special type of question where the machine needs to decide which nouns and pronouns refer to the same event or entity.

Algebraic Reasoning: A special type of question where the machine need to either accomplish algebraic operations among numbers in the passage or count entities or events in the passage.

Commonsense Reasoning: The correct option can be matched to lexical cues in the passage or inferred by synthesizing lexical cues in the passage. Often, incorrect options have matches in the passage and are confusing. Humans need commonsense knowledge as complement to the text to solve such questions. The commonsense knowledge is neither semantics or syntax of the language, nor anything that can easily be obtained from contextualized embeddings of tokens in the text.

Category	Mid	High	Total
Lexical Matching	81	69	150
Inference	13	21	34
Relation Identification	11	5	16
Summary	14	28	42
Coreference Resolution	5	2	7
Algebraic Reasoning	4	4	8
Commonsense Reasoning	20	15	35
Others	1	6	7

Table 2: Summary of question counts in each category.

Others: This category includes special questions not covered in the above categories. One example question under this category requires the machine to fill in a blank in the passage.

4 Data Categorization

We randomly sample 300 answerable test questions from RACE. RACE has two parts, one containing examination questions for middle school students and the other containing questions for high school students. We sample 150 from each part. We annotate each question with one of the reasoning skills proposed. The number of questions in each category is summarized in Table 2.

5 Experiments

5.1 Models

We focus on two model architectures.

BERT: Bidirectional Encoder Representations from Transformer is a powerful and versatile model for learning language representations (Devlin et al., 2018). BERT$_{BASE}$ and BERT$_{LARGE}$ respectively consists of a stack of 12 and 24 Transformer (Vaswani et al., 2017) encoder blocks. Each model is pretrained on large amounts of textual materials and can be finetuned on multiple choice MRC tasks. BERT achieves an overall accuracy that is close to the state-of-the-art machine performance on RACE.

DCMN: Dual Co-Matching Network (Zhang et al., 2019) uses word representations from BERT and performs dual co-matching among passage, question, and options. DCMN computes passage-aware question representations, question-aware passage representations, passage-aware option representations, and option-aware passage representations. These representations are then aggregated to perform prediction. The ensemble version of DCMN that uses word representations from BERT$_{LARGE}$ achieves the current state-of-the-art.

Model	Reproduction	Leaderboard
BERT$_{BASE}$	0.664	0.650
DCMN$_{BASE}$	0.665	-
BERT$_{LARGE}$	0.706	0.720
DCMN$_{LARGE}$	-	0.723

Table 3: Comparison of our implementation performances and leaderboard data.

5.2 Implementation Details

We experiment with BERT$_{BASE}$, BERT$_{LARGE}$, and DCMN$_{BASE}$ which uses BERT$_{BASE}$ instead of BERT$_{LARGE}$ as the encoder due to the limitation of computing resources. All models are initialized with pretrained parameters from uncased BERT and trained with max sequence length 400, learning rate 2e-5, and 16-bit training. BERT$_{BASE}$ and DCMN$_{BASE}$ are trained for 20 epochs, while BERT$_{LARGE}$ is finetuned for 5 epochs. The overall test performance is compared to the results on the leaderboard in Table 3.

Our result for BERT$_{BASE}$ is higher than the one from the leaderboard, likely because we trained for more epochs with larger sequence length. Our result for BERT$_{LARGE}$ is slightly lower than the one from the leaderboard, also likely due to differences in hyperparameter tuning. Adding dual co-matching demonstrates a little better performance than directly finetuning BERT on the multiple choice task both in our reproduction and on the leaderboard. Our reproduction is reasonably well for the purpose of investigating model performances on questions that require different types of skills.

5.3 Evaluation

We calculate the model test accuracies on different categories of questions. The results are summarized in Table 4.

All three models perform the best on lexical matching questions. Interestingly, BERT$_{BASE}$ perform slightly better than the other two models on lexical matching. Nevertheless, BERT$_{LARGE}$ and DCMN$_{BASE}$ are relatively advantageous on more complex tasks such as inference, summary, and coreference resolution. This is expected because they involve more parameters and

Category	BERT$_{BASE}$	BERT$_{LARGE}$	DCMN$_{BASE}$
Lexical Matching	**0.9200**	0.8800	0.9000
Inference	0.5294	**0.7353**	0.5882
Relation Identification	0.1875	**0.5000**	0.4375
Summary	0.6667	**0.7857**	0.6667
Coreference Resolution	0.2857	**0.8571**	0.5714
Algebraic Reasoning	**0.3750**	**0.3750**	0.1250
Commonsense Reasoning	0.3714	**0.4857**	0.4571
Others	0.7143	0.7143	0.7143

Table 4: Test accuracies on different categories of questions.

more complex mappings. However, it is unclear whether further increment in model size is a possible and efficient way for human-level performance on these questions.

Significantly, all models perform very poorly on tasks such as relation identification, algebraic reasoning, and commonsense reasoning. These types of RACE questions are usually very intuitive for humans. This shows the necessity of developing mechanisms that explicitly model these reasoning types, if we aim at human-level machine reading comprehension.

6 Discussion

We have demonstrated that existing models have much room for improvement in terms of representing a number of reasoning skills. Such improvement is necessary for human-level MRC. Much follow-up work can be done. Firstly, since BERT$_{LARGE}$ performs better than BERT$_{BASE}$ on synthesizing lexical cues, it would be meaningful to explore how further increment in the model depth may improve the performance. Secondly, since BERT$_{LARGE}$ perform badly on certain reasoning types, including relation identification and algebraic reasoning, it is necessary to explore architectures or modules that explicitly model these skills and are trained for these skills. Thirdly, datasets that focus on reasoning skills can be developed, since existing models indeed perform pretty well on lexical matching questions. Such datasets could potentially include reasoning behind decisions, so that machines can learn the mapping among different reasoning materials in a supervised and interpretable way.

References

Christopher JC Burges. 2013. Towards the machine comprehension of text: An essay. *TechReport: MSR-TR-2013-125*.

Danqi Chen, Jason Bolton, and Christopher D Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. *arXiv preprint arXiv:1606.02858*.

Danqi Chen, Christopher D. Manning, Dan Jurafsky, Percy Liang, and Luke Zettlemoyer. 2018. *Neural reading comprehension and beyond*. Ph.D. thesis.

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 arXiv:1803.05457*.

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

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

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

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

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

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

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 arXiv:1611.09268*.

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2016. Who did what: A large-scale person-centered cloze dataset. *arXiv preprint arXiv:1608.05457*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

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 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203.

Saku Sugawara, Hikaru Yokono, and Akiko Aizawa. 2017. Prerequisite skills for reading comprehension: Multi-perspective analysis of mctest datasets and systems. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

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 *Advances in neural information processing systems*, pages 5998–6008.

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