How Much *Reading* Does Reading Comprehension Require? A Critical Investigation of Popular Benchmarks

Divyansh Kaushik
Language Technologies Institute
Carnegie Mellon University
dkaushik@cs.cmu.edu

Zachary C. Lipton
Tepper School of Business
Carnegie Mellon University
zlipton@cmu.edu

Abstract

Many recent papers address *reading comprehension*, where examples consist of (*question*, *passage*, *answer*) tuples. Presumably, a model must combine information from *both* questions and passages to predict corresponding answers. However, despite intense interest in the topic, with hundreds of published papers vying for leaderboard dominance, basic questions about the difficulty of many popular benchmarks remain unanswered. In this paper, we establish sensible baselines for the bAbI, SQuAD, CBT, CNN, and WhoDidWhat datasets, finding that question- and passage-only models often perform surprisingly well. On 14 out of 20 bAbI tasks, passage-only models achieve greater than 50% accuracy, sometimes matching the full model. Interestingly, while CBT provides 20-sentence passages, only the last is needed for comparably accurate prediction. By comparison, SQuAD and CNN appear better-constructed.

1 Introduction

Recently, *reading comprehension* (RC) has emerged as a popular task, with researchers proposing various end-to-end deep learning algorithms to push the needle on a variety of benchmarks. As characterized by Hermann et al. (2015); Onishi et al. (2016), unlike prior work addressing question answering from general structured knowledge, RC requires that a model extract information from a given, unstructured passage. It's not hard to imagine how such systems could be useful. In contrast to generic text summarization, RC systems could answer targeted questions about specific documents, efficiently extracting facts and insights.

While many RC datasets have been proposed over the years (Hirschman et al., 1999; Breck et al., 2001; Peñas et al., 2011; Peñas et al., 2012; Sutcliffe et al., 2013; Richardson et al., 2013; Berant et al., 2014), more recently, larger datasets have been proposed to accommodate the data-intensiveness of deep learning. These vary both in the source and size of their corpora and in how they cast the prediction problem—as a classification task (Hill et al., 2016; Hermann et al., 2015; Onishi et al., 2016; Lai et al., 2017; Weston et al., 2016; Miller et al., 2016), span selection (Rajpurkar et al., 2016; Trischler et al., 2017), sentence retrieval (Wang et al., 2007; Yang et al., 2015), or free-form answer generation (Nguyen et al., 2016).[1] Researchers have steadily advanced on these benchmarks, proposing myriad neural network architectures aimed at attending to both questions and passages to produce answers.

In this paper, we argue that amid this rapid progress on empirical benchmarks, crucial steps are sometimes skipped. In particular, we demonstrate that the level of difficulty for several of these tasks is poorly characterized. For example, for many RC datasets, it's not reported, either in the papers introducing the datasets, or in those proposing models, how well one can perform while ignoring either the question or the passage. In other datasets, although the passage might consist of many lines of text, it's not clear how many are actually required to answer the question, e.g., the answer may always lie in the first or the last sentence.

We describe several popular RC datasets and models proposed for these tasks, analyzing their performance when provided with question-only (Q-only) or passage-only (P-only) information. We show that on many tasks, the results obtained are surprisingly strong, outperforming many base-

[1] We note several other QA datasets (Yang et al., 2015; Miller et al., 2016; Nguyen et al., 2016; Paperno et al., 2016; Clark and Etzioni, 2016; Lai et al., 2017; Trischler et al., 2017; Joshi et al., 2017) not addressed in this paper.

lines, and sometimes even surpassing the same models, supplied with both questions *and* passages.

We note that similar problems were shown for datasets in *visual question answering* by Goyal et al. (2017) and for *natural language inference* by Gururangan et al. (2018); Poliak et al. (2018); Glockner et al. (2018). Several other papers have discussed the weaknesses of various RC benchamrks (Chen et al., 2016; Lee et al., 2016). We discuss these studies in the paragraphs introducing the corresponding datasets below.

2 Datasets

In the following section, we provide context on each dataset that we investigate and then describe our process for corrupting the data as required by our question- and passage-only experiments.

CBT Hill et al. (2016) prepared a cloze-style (*fill in the blank*) RC dataset by using passages from children's books. In their dataset, each passage consists of 20 consecutive sentences, and each question is the 21st sentence with one word removed. The missing word then serves as the answer. The dataset is split into four categories of answers: Named Entities (NE), Common Nouns (CN), Verbs (V) and Prepositions (P). The training corpus contains over $37,000$ candidates and each question is associated with 10 candidates, POS-matched to the correct answer. The authors established LSTM/embedding-based Q-only baselines but did not present the results obtained by their best model using Q-only or P-only information.

CNN Hermann et al. (2015) introduced the CNN/Daily Mail datasets containing more than 1 million news articles, each associated with several highlight sentences. Also adopting the cloze-style dataset preparation, they remove an entity (answer) from a highlight (question). They anonymize all entities to ensure that models rely on information contained in the passage, vs memorizing characteristics of given entities across examples, and thus ignoring passages. On average, passages contain 26 entities, with over 500 total possible answer candidates. Chen et al. (2016) analyzed the difficulty of the CNN and Daily Mail tasks. They hand-engineered a set of eight features for each entity e (does e occur in the question, in the passage, etc.), showing that this simple classifier outperformed many earlier deep learning results.

Who-did-What Onishi et al. (2016) extracted pairs of news articles, each pair referring to the same events. Adopting the cloze-style, they remove a person's name (the answer) from the first sentence of one article (the question). A model must predict the answer based on the question, together with the other article in the pair (passage). Unlike CNN, Who-did-What does not anonymize entities. On average, each question is associated with 3.5 candidate answers. The authors removed several questions from their dataset to thwart simple strategies such as always predicting the name that occurs most (or first) in the passage.

bAbI Weston et al. (2016) presented a set of 20 tasks to help researchers identify and rectify the failings of their reading comprehension systems. Unlike the datasets discussed so far, the questions in this task are not cloze-style and are synthetically generated using templates. This restricts the diversity in clauses appearing in the passages. Further, this also restricts the dataset vocabulary to just 150 words, in contrast, CNN dataset has a vocabulary made of close to $120,000$ words. Memory Networks with adaptive memory, n-grams and non-linear matching were shown to obtain 100% accuracy on 12 out of 20 bAbI tasks. We note that Lee et al. (2016) previously identified that bAbI tasks might fall short as a measure of "AI-complete question answering", proposing two models based on *tensor product representations* that achieve 100% accuracy on many bAbI tasks.

SQuAD More recently, Rajpurkar et al. (2016) released the Stanford Question Answering Dataset (SQuAD) containing over $100,000$ crowd-sourced questions addressing 536 passages. Each question is associated with a paragraph (passage) extracted from an article. These passages are shorter than those in CNN and Who-did-What datasets. Models choose answers by selecting (varying-length) spans from these passages.

Generating Corrupt Data To void any information in either the questions or the passages, while otherwise leaving each architecture intact, we create corrupted versions of each dataset by assigning either questions randomly, while preserving the correspondence between passage and answer, or by randomizing the passage. For

tasks where question-answering requires selecting spans or candidates from the passage, we create passages that contain the candidates in random locations but otherwise consist of random gibberish.

3 Models

In our investigations of the various RC benchmarks, we rely upon the following three recently-proposed models: key-value memory networks, gated attention readers, and QA nets. Although space constraints preclude a full discussion of each architecture, we provide references to the source papers and briefly discuss any implementation decisions necessary to reproduce our results.

Key-Value Memory Networks We implement a Key-Value Memory Network (KV-MemNet) (Miller et al., 2016), applying it to bAbI and CBT. KV-MemNets are based on Memory Networks (Sukhbaatar et al., 2015), shown to perform well on both datasets. For bAbI tasks, the keys and values both encode the passage as a bag-of-words (BoW). For CBT, the key is a BoW-encoded 5-word window surrounding a candidate answer and the value is the candidate itself. We fixed the number of hops to 3 and the embedding size to 128.

Gated Attention Reader Introduced by Dhingra et al. (2017), the Gated Attention Reader (GAR)[2] performs multiple hops over a passage, like MemNets. The word representations are refined over each hop and are mapped by an attention-sum module (Kadlec et al., 2016) to a probability distribution over the candidate answer set in the last hop. The model nearly matches best-reported results on many cloze-style RC datasets, and thus we apply it to *Who-did-What, CNN, CBT-NE* and *CBT-CN*.

QA Net Recently introduced by (Yu et al., 2018), the QA-Net[3] was recently demonstrated to outperform all previous models on the SQuAD dataset[4]. Passages and questions are passed as input to separate encoders consisting of depth-wise separable convolutions and global self-attention. This is followed by a passage-question attention layer, followed by stacked encoders. The outputs from these encoders are used to predict an answer span inside the passage.

4 Experimental Results

bAbI tasks Table 1 shows the results obtained by a Key-Value Memory Network on bAbI tasks by nullifying the information present in either questions or passages. On tasks $2, 7, 13$ and 20, P-only models obtain over 80% accuracy with questions randomly assigned. Moreover, on tasks 3, 13, 16, and 20, P-only models match performance of those trained on the full dataset. On task 18, Q-only models achieve an accuracy of 91%, nearly matching the best performance of 93% achieved by the full model. These results show that some of bAbI tasks are easier than one might think.

Children's Books Test On the NE and CN CBT tasks, Q-only KV-MemNets obtain an accuracy close to the *full* accuracy and on the Verbs (V) and Prepositions (P) tasks, Q-only models outperform the full model (Table 2). Q-only Gated attention readers reach accuracy of 50.6% and 54% on Named Entities (NE) and Common Nouns (CN) tasks, respectively, while P-only models reach accuracies of 40.8% and 36.7%, respectively. We note that our models can outperform 16 of the 19 reported results on the NE task in Hill et al. (2016) using Q-only information. Table 3 shows that if we make use of just last sentence instead of all 20 sentences in the passage, our sentence memory based KV-MemNet achieve comparable or better performance *w.r.t* the *full* model on most subtasks.

CNN Table 2, shows the performance of Gated Attention Reader on the CNN dataset. Q-only and P-only models obtained 25.6% and 38.3% accuracies respectively, compared to 77.8% on the true dataset. This drop in accuracy could be due to the anonymization of entities which prevents models from building entity-specific information. Notwithstanding the deficiencies noted by Chen et al. (2016), we found that out CNN, out all the cloze-style RC datasets that we evaluated, appears to be the most carefully designed.

Who-did-What P-only models achieve greater than 50% accuracy in both the strict and relaxed setting, reaching within 15% of the accuracy of the *full* model in the strict setting. Q-only models also achieve 50% accuracy on the relaxed setting while achieving an accuracy of 41.8% on the strict setting. Our P-only model also outperforms all the

[2] https://github.com/bdhingra/ga-reader

[3] We use the implementation available at https://github.com/NLPLearn/QANet

[4] At the time of publication, an ensemble of QA-Net models was at the top of the leader board. A single QA-Net was ranked 4th.

	bAbI Tasks 1-10									
Dataset	1	2	3	4	5	6	7	8	9	10
True dataset	**100%**	**100%**	39%	**100%**	**99%**	**100%**	**94%**	**97%**	**99%**	**98%**
Question only	18%	17%	22%	22%	34%	50%	48%	34%	64%	44%
Passage only	53%	86%	**60%**	59%	31%	48%	85%	79%	63%	47%
$\Delta(min)$	−47	−14	+21	−41	−65	−52	−9	−18	−35	−51
	bAbI Tasks 11-20									
	11	12	13	14	15	16	17	18	19	20
True dataset	**94%**	**100%**	**94%**	**96%**	**100%**	**48%**	**57%**	**93%**	**30%**	**100%**
Question only	17%	15%	18%	18%	34%	26%	48%	91%	10%	70%
Passage only	71%	74%	**94%**	50%	64%	**47%**	48%	53%	21%	**100%**
$\Delta(min)$	−23	−26	0	−46	−36	−1	−9	−2	−9	0

Table 1: Accuracy on bAbI tasks using our implementation of the Key-Value Memory Networks

Task	Full	Q-only	P-only	$\Delta(min)$
Key-Value Memory Networks				
CBT-NE	**35.0%**	29.1%	24.1%	−5.9
CBT-CN	**37.6%**	32.4%	24.4%	−5.2
CBT-V	52.5%	**55.7%**	36.0%	+3.2
CBT-P	55.2%	**56.9%**	30.1%	+1.7
Gated Attention Reader				
CBT-NE	**74.9%**	50.6%	40.8%	−17.5
CBT-CN	**70.7%**	54.0%	36.7%	−16.7
CNN	**77.8%**	25.6%	38.3%	−39.5
WdW	**67.0%**	41.8%	52.2%	−14.8
WdW-R	**69.1%**	50.0%	50.6%	−15.6

Table 2: Accuracy on various datasets using KV-MemNets (window memory) and GARs

Task	Complete passage	Last sentence
CBT-NE	22.6%	**22.8%**
CBT-CN	**31.6%**	24.8%
CBT-V	**48.8%**	45.0%
CBT-P	34.1%	**37.9%**

Table 3: Accuracy on CBT tasks using KV-MemNets (sentence memory) varying passage size.

suppressed baselines and 5 additional baselines reported by Onishi et al. (2016). We suspect that the models memorize attributes of specific entities, justifying the entity-anonymization used by Hermann et al. (2015) to construct the CNN dataset.

Metric	Full	Q-only	P-only	$\Delta(min)$
EM	**70.7%**	0.6%	10.9%	−59.8
F1	**79.1%**	4.0%	14.8%	−64.3

Table 4: Performance of QANet on SQuAD

SQuAD Our results suggest that SQuAD is an unusually carefully-designed and challenging RC task. The span selection mode of answering requires that models consider the passage thus the abysmal performance of the Q-only QANet (Table 4). Since SQuAD requires answering by span selection, we construct Q-only variants here by placing answers from all relevant questions in random order, filling the gaps with random words. Moreover, Q-only and P-only models achieve F1 scores of only 4% and 14.8% resp. (Table 4), significantly lower than 79.1 on the proper task.

5 Discussion

We briefly discuss our findings, offer some guiding principles for evaluating new benchmarks and algorithms, and speculate on why some of these problems may have gone under the radar. Our goal is not to blame the creators of past datasets but instead to support the community by offering practical guidance for future researchers.

Provide rigorous RC baselines Published RC datasets should contain reasonable baselines that characterize the difficulty of the task, and specifically, the extent to which questions and passages are essential. Moreover, follow-up papers reporting improvements ought to report performance both on the full task and variations omitting questions and passages. While many proposed technical innovations purportedly work by better matching up information in questions and passages, absent these baselines one cannot tell whether gains come for the claimed reason or if the models just do a better job of passage classification (disregarding questions).

Test that *full context* is essential Even on tasks where both questions and passages are required, problems might appear harder than they really are. On first glance the the length-20 passages in CBT, might suggest that success requires reasoning over all 20 sentences to identify the correct answer to each question. However, it turns out that for some models, comparable performance can be achieved by considering only the last sentence. We recommend that researchers provide reasonable ablations to characterize the amount of context that each model truly requires.

Caution with cloze-style RC datasets We note that cloze-style datasets are often created programatically. Thus it's possible for a dataset to be produced, published, and incorporated into many downstream studies, all without many person-hours spent manually inspecting the data. We speculate that, as a result, these datasets tend be subject to less contemplation of what's involved in answering these questions and are therefore especially susceptible to the sorts of overlooked weaknesses described in our study.

A note on publishing incentives We express some concern that the recommended experimental rigor might cut against current publishing incentives. We speculate that papers introducing datasets may be more likely to be accepted at conferences by omitting unfavorable ablations than by including them. Moreover, with reviewers often demanding *architectural novelty*, methods papers may find an easier path to acceptance by providing unsubstantiated stories about the *reasons* why a given architecture works than by providing rigorous ablation studies stripping out spurious explanations and unnecessary model components. For more general discussions of misaligned incentives and empirical rigor in machine learning research, we point the interested reader to Lipton and Steinhardt (2018) and Sculley et al. (2018).

References

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 *Empirical Methods in Natural Language Processing (EMNLP)*.

Eric Breck, Marc Light, Gideon Mann, Ellen Riloff, Brianne Brown, and Pranav Anand. 2001. Looking under the hood: Tools for diagnosing your question answering engine. In *Association for Computational Linguistics (ACL) Workshop on Open-Domain Question Answering*.

Danqi Chen, Jason Bolton, and Christopher D Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. In *Association for Computational Linguistics (ACL)*.

Peter Clark and Oren Etzioni. 2016. My computer is an honor student but how intelligent is it? standardized tests as a measure of ai. *AI Magazine*, 37(1):5–12.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2017. Gated-attention readers for text comprehension. In *Association for Computational Linguistics (ACL)*.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking nli systems with sentences that require simple lexical inferences. In *Association for Computational Linguistics (ACL)*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Computer Vision and Pattern Recognition (CVPR)*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies*.

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 (NIPS)*.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The Goldilocks principle: Reading children's books with explicit memory representations. In *International Conference on Learning Representations (ICLR)*.

Lynette Hirschman, Marc Light, Eric Breck, and John D Burger. 1999. Deep read: A reading comprehension system. In *Association for Computational Linguistics on Computational Linguistics (ACL)*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Association for Computational Linguistics (ACL)*.

Rudolf Kadlec, Martin Schmid, Ondřej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. In *Association for Computational Linguistics (ACL)*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Moontae Lee, Xiaodong He, Wen-tau Yih, Jianfeng Gao, Li Deng, and Paul Smolensky. 2016. Reasoning in vector space: An exploratory study of question answering. In *International Conference on Learning Representations (ICLR)*.

Zachary C Lipton and Jacob Steinhardt. 2018. Troubling trends in machine learning scholarship. In *International Conference on Machine Learning (ICML) Machine Learning Debates Workshop*.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernandez. 2016. The lambada dataset: Word prediction requiring a broad discourse context. In *Association for Computational Linguistics (ACL)*.

Anselmo Peñas, Eduard Hovy, Pamela Forner, Álvaro Rodrigo, Richard Sutcliffe, Caroline Sporleder, Corina Forascu, Yassine Benajiba, and Petya Osenova. 2012. Overview of qa4mre at clef 2012: Question answering for machine reading evaluation.

Anselmo Peñas, Eduard Hovy, Pamela Forner, Álvaro Rodrigo, Richard Sutcliffe, Corina Forascu, and Caroline Sporleder. 2011. Overview of qa4mre at clef 2011: Question answering for machine reading evaluation.

A. Poliak, J. Naradowsky, A. Haldar, R. Rudinger, and B. Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Joint Conference on Lexical and Computational Semantics (*SEM)*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Empirical Methods in Natural Language Processing (EMNLP)*.

D Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. 2018. Winner's curse? on pace, progress, and empirical rigor. In *International Conference on Learning Representations (ICLR) Workshop Track*.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems (NIPS)*.

Richard Sutcliffe, Anselmo Peñas, Eduard Hovy, Pamela Forner, Álvaro Rodrigo, Corina Forascu, Yassine Benajiba, and Petya Osenova. 2013. Overview of qa4mre main task at clef 2013.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. Newsqa: A machine comprehension dataset. In *Workshop on Representation Learning for NLP (RepL4NLP)*.

Mengqiu Wang, Noah A Smith, and Teruko Mitamura. 2007. What is the jeopardy model? a quasi-synchronous grammar for qa. In *Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2016. Towards ai-complete question answering: A set of prerequisite toy tasks. In *International Conference on Learning Representations (ICLR)*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Empirical Methods in Natural Language Processing (EMNLP)*.

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 *International Conference on Learning Representations (ICLR)*.