

# ORB: An Open Reading Benchmark for Comprehensive Evaluation of Machine Reading Comprehension

Dheeru Dua<sup>♣</sup>, Ananth Gottumukkala<sup>♣</sup>, Alon Talmor<sup>♡</sup>,  
Sameer Singh<sup>♣</sup>, and Matt Gardner<sup>♣</sup>

<sup>♣</sup>University of California, Irvine, USA

<sup>♡</sup>Allen Institute for Artificial Intelligence, Seattle, Washington, USA

<sup>♣</sup>Allen Institute for Artificial Intelligence, Irvine, California, USA

ddua@uci.edu

## Abstract

Reading comprehension is one of the crucial tasks for furthering research in natural language understanding. A lot of diverse reading comprehension datasets have recently been introduced to study various phenomena in natural language, ranging from simple paraphrase matching and entity typing to entity tracking and understanding the implications of the context. Given the availability of many such datasets, comprehensive and reliable evaluation is tedious and time-consuming for researchers working on this problem. We present an evaluation server, **ORB**, that reports performance on seven diverse reading comprehension datasets, encouraging and facilitating testing a single model’s capability in understanding a wide variety of reading phenomena. The evaluation server places no restrictions on how models are trained, so it is a suitable test bed for exploring training paradigms and representation learning for general reading facility. As more suitable datasets are released, they will be added to the evaluation server. We also collect and include synthetic augmentations for these datasets, testing how well models can handle out-of-domain questions.

## 1 Introduction

Research in reading comprehension, the task of answering questions about a given passage of text, has seen a huge surge of interest in recent years, with many large datasets introduced targeting various aspects of reading (Rajpurkar et al., 2016; Dua et al., 2019; Dasigi et al., 2019; Lin et al., 2019). However, as the number of datasets increases, evaluation on all of them becomes challenging, encouraging researchers to overfit to the biases of a single dataset. Recent research, including MultiQA (Talmor and Berant, 2019) and the MRQA workshop shared task, aim to facilitate training and evaluating on several reading comprehension datasets at

the same time. To further aid in this direction, we present an evaluation server that can test a single model across many different datasets, including on their hidden test sets in some cases. We focus on datasets where the core problem is natural language understanding, not information retrieval; models are given a single passage of text and a single question and are required to produce an answer.

As our goal is to provide a broad suite of questions that test a single model’s reading ability, we additionally provide synthetic augmentations to some of the datasets in our evaluation server. Several recent papers have proposed question transformations that result in out-of-distribution test examples, helping to judge the generalization capability of reading models (Ribeiro et al., 2018, 2019; Zhu et al., 2019). We collect the best of these, add some of our own, and keep those that generate reasonable and challenging questions. We believe this strategy of evaluating on many datasets, including distribution-shifted synthetic examples, will lead the field towards more robust and comprehensive reading comprehension models.

Code for the evaluation server, including a script to run it on the dev sets of these datasets and a leaderboard showing results on their hidden tests, can be found at <https://leaderboard.allenai.org/orb>

## 2 Datasets

We selected seven existing datasets that target various complex linguistic phenomena such as coreference resolution, entity and event detection, etc., capabilities which are desirable when testing a model for reading comprehension. We chose datasets that adhere to two main properties: First, we exclude from consideration any multiple choice dataset, as these typically require very different model architectures, and they often have biases in how the dis-

tractor choices are generated. Second, we require that the dataset be originally designed for answering isolated questions over a single, given passage of text. We are focused on evaluating *reading* performance, not *retrieval*; reading a single passage of text is far from solved, so we do not complicate things by adding in retrieval, conversation state, or other similar complexities.

It is our intent to add to the evaluation server any high-quality reading comprehension dataset that is released in the future that matches these restrictions.

We now briefly describe the datasets that we include in the initial release of ORB. Table 1 gives summary statistics of these datasets. Except where noted, we include both the development and test sets (including hidden test sets) in our evaluation server for all datasets.

**SQuAD** (Rajpurkar et al., 2016) requires a model to perform lexical matching between the context and the question to predict the answer. This dataset provides avenues to learn predicate-argument structure and multi-sentence reasoning to some extent. It was collected by asking crowd-workers to create question-answer pairs from Wikipedia articles such that the answer is a single-span in the context. The dataset was later updated to include unanswerable questions (Rajpurkar et al., 2018), giving a harder question set without as many reasoning shortcuts. We include only the development sets of SQuAD 1.1 and SQuAD 2.0 in our evaluation server.

**DuoRC** (Saha et al., 2018) tests if the model can generalize to answering semantically similar but syntactically different paraphrased questions. The questions are created on movie summaries obtained from two sources, Wikipedia and IMDB. The crowd-workers formalized questions based on Wikipedia contexts and in turn answered them based on the IMDB context. This ensured that the model will not rely solely on lexical matching, but rather utilize semantic understanding. The answer can be either a single-span from context or free form text written by the annotator.

**Quoref** (Dasigi et al., 2019) focuses on understanding coreference resolution, a challenging aspect of natural language understanding. It helps gauge how a model can handle ambiguous entity and event resolution to answer a question correctly. This dataset was created by asking crowd workers to write questions and multi-span answers from

Wikipedia articles that centered around pronouns in the context.

**DROP** (Dua et al., 2019) attempts to force models to have a more comprehensive understanding of a paragraph, by constructing questions that query many parts of the paragraph at the same time. These questions involve reasoning operations that are mainly rudimentary mathematical skills such as addition, subtraction, maximum, minimum, etc. To perform well on this dataset a model needs to locate multiple spans associated with questions in the context and perform a set of operations in a hierarchical or sequential manner to obtain the answer. The answer can be either a set of spans from the context, a number or a date.

**ROPES** (Lin et al., 2019) centers around understanding the implications of a passage of text, particularly dealing with the language of causes and effects. A system is given a background passage, perhaps describing the effects of deforestation on local climate and ecosystems, and a grounded situation involving the knowledge in the background passage, such as, *City A has more trees than City B*. The questions then require grounding the effects described in the background, perhaps querying which city would more likely have greater ecological diversity. This dataset can be helpful in understanding how to apply the knowledge contained in a passage of text to a new situation.

**NewsQA** (Trischler et al., 2017) dataset focuses on paraphrased questions with predicate-argument structure understanding. To some extent it is similar to DuoRC, however the examples are collected from news articles and offers diverse linguistic structures. This crowd-sourced dataset was created by asking annotators to write questions from CNN/DailyMail articles as context.

**NarrativeQA** (Kočiský et al., 2018) focuses on understanding temporal reasoning among various events that happen in a given movie plot. It also tests the models ability to “hop” between various parts of the context and not rely solely on sequential reasoning. The dataset was constructed by aligning books from Gutenberg<sup>1</sup> with the summaries of their movie adaptations from various web resources. The crowd workers were asked to create complex questions about characters, narratives, events etc. from summaries and typically can be answered

<sup>1</sup><http://www.gutenberg.org/>

Dataset	Dev Size	Test Size	Context Length (Avg)	Answer Length (Avg)
SQuAD1.1	10,570	-	123.7	4.0
SQuAD2.0	10,570	-	127.5	4.2
DuoRC	12,233	13,449	1113.6	2.8
Quoref	2,418	2,537	348.2	2.7
DROP	9,536	9,622	195.1	1.5
ROPES	1,204	1,015	177.1	1.2
NewsQA	5,166	5,126	711.3	5.1
NarrativeQA	3,443	10,557	567.9	4.7

Table 1: Dataset Statistics

from summaries. In addition, crowd workers were required to provide answers that do not have high overlap with the context. In accordance with our format, we only use the version with the summaries as context in our evaluation server.

### 3 Synthetic Augmentations

Prior works (Jia and Liang, 2017) have shown that RC models are brittle to minor perturbations in original dataset. Hence, to test the model’s ability to generalize to out-of-domain syntactic structures and be logically consistent in its answers, we automatically generate questions based on various heuristics. These heuristics fall in two broad categories.

1. The question is paraphrased to a minimal extent to create new syntactic structures, keeping the semantics of the question largely intact and without making any changes to the original context and answer.
2. The predicate-argument structures of a given question-answer pair are leveraged to create new WH-questions based on the object in the question instead of the subject. This rule-based method, adopted from (Ribeiro et al., 2019), changes the question and answer keeping the context fixed.

We use five augmentation techniques, where the first four techniques fall into the first category and the last technique falls into the second category.

**Invert Choice** transforms a binary choice question by changing the order in which the choices are presented, keeping the answer the same.

**More Wrong Choice** transforms a binary choice question by substituting the wrong choice in the question with another wrong choice from the passage.

Dataset	IC	MWC	Imp	No-Ans	SEARs
NewsQA	0	0	501	347	16009
QuoRef	0	0	79	385	11759
DROP	1377	457	113	284	16382
SQuAD	16	0	875	594	28188
ROPES	637	119	0	201	2909
DuoRC	22	0	2706	-	45020

Table 2: Yields of augmented datasets

**No Answer** substitutes a name in the question for a different name from the passage to create with high probability a new question with no answer.

**SEARs** creates minimal changes in word selection or grammar while maintaining the original meaning of the question according to the rules described by Ribeiro et al. (2018).

**Implication** creates a new question-answer pair, where the object of the original question is replaced with the answer directly resulting in creation of a new WH-question where the answer is now the object of the original question. These transformations are performed based on rules described by Ribeiro et al. (2019).

We attempted all the above augmentation techniques on all the datasets (except NarrativeQA where entity and event tracking is complex and these simple transformations can lead to a high number of false positives). Table 2 shows the number of augmentations generated by each augmentation technique-dataset pair. A few sample augmentations are shown in Table 3.

After generating all the augmented datasets, we manually identified the augmentation technique-dataset pairs which led to high-quality augmentations. We sample 50 questions from each augmented dataset and record whether they satisfy the three criteria given below.

1. Is the question understandable, with little to no grammatical errors?
2. Is the question semantically correct?
3. Is the new answer the correct answer for the new question?

Table 4 shows the number of high-quality questions generated for each dataset. We keep the augmentation technique-dataset pairs where at least 90% of the question-answer pairs satisfy the above three criteria. We further test the performance of these augmentations (Section 4) on a BERT (Devlin et al., 2019b) based model to establish if the dataset has a sufficiently different question distribu-

Template Type	Context (truncated)	Original QA Pair	Generated QA Pair
<b>Invert Choice</b>	... before halftime thanks to a David Akers 32-yard field goal, giving Detroit a 17-14 edge ... in the third, Washington was able to equalize with John Potter making his first career field goal from 43 yards out ... in the fourth, Detroit took the lead again, this time by way of Akers hitting a 28-yard field goal, giving Detroit a 20-17 lead...	<b>Q:</b> Which player scored more field goals, David Akers or John Potter? <b>A:</b> David Akers	<b>Q:</b> Which player scored more field goals, John Potter or David Akers? <b>A:</b> David Akers
<b>More Wrong Choice</b>	The first issue in 1942 consisted of denominations of 1, 5, 10 and 50 centavos and 1, 5, and 10 Pesos. ... 1944 ushered in a 100 Peso note and soon after an inflationary 500 Pesos note. In 1945, the Japanese issued a 1,000 Pesos note...	<b>Q:</b> Which year ushered in the largest Pesos note, 1944 or 1945? <b>A:</b> 1945	<b>Q:</b> Which year ushered in the largest Pesos note, 1942 or 1945? <b>A:</b> 1945
<b>Implication</b>	... In 1562, naval officer Jean Ribault led an expedition that explored Florida and the present-day Southeastern U.S., and founded the outpost of Charlesfort on Parris Island, South Carolina...	<b>Q:</b> When did Ribault first establish a settlement in South Carolina? <b>A:</b> 1562	<b>Q:</b> Who established a settlement in South Carolina in 1562? <b>A:</b> Ribault
<b>No Answer</b>	From 1975, Flavin installed permanent works in Europe and the United States, including ... the Union Bank of Switzerland, Bern (1996). ... The 1930s church was designed by Giovanni Muzio...	<b>Q:</b> Which permanent works did Flavin install in 1996? <b>A:</b> Union Bank of Switzerland, Bern	<b>Q:</b> Which permanent works did Giovanni Muzio install in 1996? <b>A:</b> No Answer
<b>SEARs</b>	... Dhul-Nun al-Misri and Ibn Wahshiyya were the first historians to study hieroglyphs, by comparing them to the contemporary Coptic language used by Coptic priests in their time...	<b>Q:</b> What did historians compare to the Coptic language? <b>A:</b> hieroglyphs	<b>Q:</b> What'd historians compare to the Coptic language? <b>A:</b> hieroglyphs

Table 3: Examples of generated augmentations with various templates.

Dataset	IC	MWC	Imp	No-Ans	SEARs
NewsQA	-	-	47	47	50
QuoRef	-	-	45	48	50
DROP	46	42	36	48	50
SQuAD	15/16	-	47	48	50
ROPES	48	36	-	11	50
DuoRC	18/22	-	47	-	50

Table 4: Quality of augmented datasets (# of good questions out of 50 sampled)

tion from the original and has enough independent value to be incorporated into the evaluation server.

## 4 Experiments

### 4.1 Model

We train a numerically-aware BERT-based model<sup>2</sup> (NABERT) on all the seven datasets and test its performance on existing datasets and synthetic augmentations. NABERT is a BERT based model with the ability to perform discrete operations like counting, addition, subtraction etc. We added support for “impossible” answers in the existing NABERT architecture by extending the answer type predictor which classifies the type of reasoning involved given a question into one of the following five categories: *number*, *span*, *date*, *count*, *impossible*. All the hyper-parameter settings were kept the same.

<sup>2</sup><https://github.com/raylin1000/drop-bert>

We noticed *catastrophic forgetting* on randomly sampling a minibatch for training, from all the datasets pooled together. To alleviate this problem, we sampled uniformly from each dataset in the beginning and then switched to sampling in proportion to the size of each dataset towards the end of the epoch (Stickland and Murray, 2019). This helped improve the performance on several dataset by 3-4% in EM, however, there is still a lot of room for improvement on this front. We also tried a simple BERT model (Devlin et al., 2019a) and MultiQA (Talmor and Berant, 2019) but NABERT gave the best results on the seven development sets.

In case of DuoRC and NarrativeQA, some answers are free-form human generated and do not have an exact overlap with the context. However, the NABERT model is trained to predict a span’s start and end indices in the context. So for answers, which are not exact spans from the context we pick a span which has the highest ROUGE-L with the gold answer to serve as labels for training. However, for evaluation we use the original gold answer and not the extracted passage span for evaluating the model’s performance.

### 4.2 Existing Dataset Performance

Table 5 shows the result of evaluating on all of the development and test sets using our evalua-

Dataset	Dev		Test	
	EM	F <sub>1</sub>	EM	F <sub>1</sub>
NewsQA	29.34	45.40	29.69	46.19
Quoref	34.49	42.65	30.13	38.39
DROP	19.09	23.16	17.69	21.87
SQuAD 1.1	68.03	78.55	-	-
SQuAD 2.0	33.70	39.17	-	-
ROPES	40.03	49.07	47.96	56.06
DuoRC	25.65	34.28	23.44	31.73

  

Narrative QA	BLEU-1	BLEU-4	METEOR	ROUGE-L (F1)
Dev Set	0.17	0.021	0.33	0.52
Test Set	0.16	0.019	0.33	0.53

Table 5: Performance on baseline BERT model on different datasets

tion server. We chose the official metrics adopted by the individual datasets to evaluate the performance of our baseline model. As can be seen in the table, the results are quite poor, significantly below single-dataset state-of-the-art on all datasets. The training of our initial baseline appears to be dominated by SQuAD 1.1, or perhaps SQuAD 1.1 mainly tests reasoning that is common to all of the other datasets. Significant research is required to build reading systems and develop training regimes that are general enough to handle multiple reading comprehension datasets at the same time, even when all of the datasets are seen at training time.

### 4.3 Synthetic Augmentations

Table 6 shows the performance of the baseline model on various development sets and heuristically generated questions. The **More Wrong Choice** augmentation is omitted since a high enough quality and/or yield of questions could not be ensured for any of the datasets. When evaluated on out-of-domain linguistic structures, performance drops significantly for some augmentation-dataset pairs but only marginally for others. For questions generated by the **Invert Choice** augmentation, the model struggles to grasp the correct reasoning behind two answer options like *Art Euphoric* or *Trescott Street* and changes the prediction when the choices are flipped. However, relative to the dev set performances on the original datasets, the performance drop is almost nonexistent. For the **SEARs** based augmentation the generated linguistic variations are close to in-domain syntactic structure so we do not see much performance drop in most of the datasets except for ROPES and

NewsQA. The **Implication** style questions create a large performance drop for NewsQA and SQuAD while having a performance boost for DuoRC. Finally, the **No-Ans** type questions have the worst performance across board for all datasets.

## 5 Related Work

### Generalization and multi-dataset evaluation

Recently there has been some work aimed at exploring the relation and differences between multiple reading comprehension datasets.

MULTIQA (Talmor and Berant, 2019) investigates over ten RC datasets, training on one or more source RC datasets, and evaluating generalization, as well as transfer to a target RC dataset. This work analyzes the factors that contribute to generalization, and shows that training on a source RC dataset and transferring to a target dataset substantially improves performance. MultiQA also provides a single format including a model and infrastructure for training and comparing question answering datasets. We provide no training mechanism, instead focusing on very simple evaluation that is compatible with any training regime, including evaluating on hidden test sets.

MRQA 19, the Machine Reading for Question Answering workshop, introduced a shared task, which tests whether existing machine reading comprehension systems can generalize beyond the datasets on which they were trained. The task provides six large-scale datasets for training, and evaluates generalization to ten different hidden test datasets. However these datasets were modified from their original version, and context was limited to 800 tokens. In addition this shared task only tests for generalization with no intra-domain evaluation. In contrast, our evaluation server simply provides a single-model evaluation on many different datasets, with no prescriptions about training regimes.

**NLP evaluation benchmarks** The General Language Understanding Evaluation benchmark or GLUE (Wang et al., 2018) is a tool for evaluating and analyzing the performance of models across a diverse range of existing NLU tasks. A newer version, Super-GLUE (Wang et al., 2019) is styled after GLUE with a new set of more difficult language understanding tasks. In this line of work another standard toolkit for evaluating the quality of universal sentence representations is SENTENCEVAL (Conneau and Kiela, 2018). Similar to GLUE,

	Dev		IC		Imp		No-Ans		SEARs	
	EM	F <sub>1</sub>	EM	F <sub>1</sub>	EM	F <sub>1</sub>	EM	F <sub>1</sub>	EM	F <sub>1</sub>
NewsQA	29.34	45.40	-	-	23.35	34.36	0.02	0.02	21.34	33.33
QuoRef	34.49	42.65	-	-	32.91	44.84	0.0	0.0	34.84	42.11
DROP	19.09	23.16	40.23	48.03	-	-	0.0	0.0	16.97	21.65
SQuAD	68.03	78.55	56.25	64.58	46.74	57.97	0.0	0.0	56.53	71.25
ROPES	40.03	49.07	24.08	31.74	-	-	-	-	14.05	19.12
DuoRC	25.65	34.28	27.27	34.19	30.30	35.23	-	-	21.51	28.85

Template Type	Answered Incorrectly	Answered Correctly
Invert Choice	<b>Original:</b> Which art gallery was founded first, Art Euphoric or Trescott Street? <b>Generated:</b> Which art gallery was founded first, Trescott Street or Art Euphoric?	<b>Original:</b> Who scored more field goals, Nate Kaeding or Dan Carpenter? <b>Generated:</b> Who scored more field goals, Dan Carpenter or Nate Kaeding?
Implication	<b>Original:</b> When did the Huguenots secure the right to own land in the Baronies? <b>Generated:</b> Who secured the right to own land in baronies in 1697?	<b>Original:</b> When did Henry issue the Edict of Nantes? <b>Generated:</b> What did Henry issue in 1598?
SEARs	<b>Original:</b> What was the theme of Super Bowl 50? <b>Generated:</b> So what was the theme of Super Bowl 50?	<b>Original:</b> Who won Super Bowl 50? <b>Generated:</b> So who won Super Bowl 50?

Table 6: Quantitative and qualitative analysis of generated augmentations. We only show performance for high yield and high-quality augmentations.

SENTEVAL also encompasses a variety of tasks, including binary and multi-class classification, natural language inference and sentence similarity. We differ from GLUE and SENTEVAL by focusing on reading comprehension tasks, and only evaluating a single model on all datasets, instead of allowing the model to be tuned to each dataset separately.

### Evaluation Platforms and Competitions in NLP

The use of online evaluation platform with private test labels has been exercised by various leaderboards on Kaggle and CodaLab, as well as shared tasks at the SemEval and CoNLL conferences.

Additional benchmarks such as PARLAI (Miller et al., 2017) and BABI (Weston et al., 2016) proposed a hierarchy of tasks towards building question answering and reasoning models and language understanding. However these frameworks do not include a standardized evaluation suite for system performance nor do they offer a wide set of reading comprehension tasks.

## 6 Conclusion

We have presented ORB, an open reading benchmark designed to be a comprehensive test of reading comprehension systems, in terms of their gen-

eralizability, understanding of various natural language phenomenon, capability to make consistent predictions, and ability to handle out-of-domain questions. This benchmark will grow over time as more interesting and useful reading comprehension datasets are released. We hope that this benchmark will help drive research on general reading systems.

## References

- A. Conneau and D. Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *Association for Computational Linguistics (ACL)*.
- Pradeep Dasigi, Nelson Liu, Ana Marasovic, Noah Smith, and Matt Gardner. 2019. QuoRef: A reading comprehension dataset with questions requiring coreferential reasoning. In *EMNLP*.
- J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019a. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Association for Computational Linguistics (NAACL)*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. Bert: Pre-training of deep bidirectional transformers for language understanding. *North American Association for Computational Linguistics (NAACL)*.

- D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *North American Association for Computational Linguistics (NAACL)*.
- R. Jia and P. Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Empirical Methods in Natural Language Processing (EMNLP)*.
- T. Kočiský, J. Schwarz, P. Blunsom, C. Dyer, K. Hermann, G. Melis, and E. Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics (TACL)*, 6:317–328.
- Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019. Reasoning over paragraph effects in situations. *arXiv preprint arXiv:1908.05852*.
- A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. 2017. Parlai: A dialog research software platform. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–84.
- P. Rajpurkar, R. Jia, and P. Liang. 2018. Know what you don’t know: Unanswerable questions for SQuAD. In *Association for Computational Linguistics (ACL)*.
- P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Empirical Methods in Natural Language Processing (EMNLP)*.
- Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 6174–6184.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865.
- A. Saha, R. Aralikkatte, M. Khapra, and K. Sankaranarayanan. 2018. Duorc: Towards complex language understanding with paraphrased reading comprehension. In *Association for Computational Linguistics (ACL)*.
- Asa Cooper Stickland and Iain Murray. 2019. BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5986–5995, Long Beach, California, USA. PMLR.
- Alon Talmor and Jonathan Berant. 2019. Multiqa: An empirical investigation of generalization and transfer in reading comprehension. *57th Annual Conference of the Association for Computational Linguistics (ACL)*, abs/1905.13453.
- A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman. 2017. NewsQA: A machine comprehension dataset. In *Workshop on Representation Learning for NLP*.
- A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.
- J. Weston, A. Bordes, S. Chopra, and T. Mikolov. 2016. Towards AI-complete question answering: A set of prerequisite toy tasks. *ICLR*.
- Haichao Zhu, Li Dong, Furu Wei, Wenhui Wang, Bing Qin, and Ting Liu. 2019. Learning to ask unanswerable questions for machine reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4238–4248, Florence, Italy. Association for Computational Linguistics.