ScholarlyRead: A New Dataset for Scientific Article Reading Comprehension

Tanik Saikh, Asif Ekbal, Pushpak Bhattacharyya
Indian Institute of Technology Patna
Bihta, Patna, Bihar-801106, India
{tanik.srf17, asif, pb}@iitp.ac.in

Abstract

We present ScholarlyRead, span-of-word-based scholarly articles' Reading Comprehension (RC) dataset with approximately 10K manually checked passage-question-answer instances. ScholarlyRead was constructed in semi-automatic way. We consider the articles from two popular journals of a reputed publishing house. Firstly, we generate questions from these articles in an automatic way. Generated questions are then manually checked by the human annotators. We propose a baseline model based on Bi-Directional Attention Flow (BiDAF) network that yields the F1 score of 37.31%. The framework would be useful for building Question-Answering (QA) systems on scientific articles.

1. Introduction

Resurgence of Artificial Intelligence (AI) and various deep learning techniques have solved many complicated tasks successfully. It is also possible to learn a machine to comprehend a document. The machine will be able to yield answers to questions based on that particular document. This task is typically called as Machine Reading Comprehension (MRC)/Reading Comprehension (RC) (Hermann et al., 2015). The accompanying document (that remains the information, essential for answering the questions) often termed as Context. Comprehension is the capability to understand something. Learning machine to understand human language documents is one of the most elusive and well established challenges in AI. Prior works and datasets on MRC mostly concern with the general domain datasets of news articles and/or elementary school-level storybooks and/or on the articles contained in Wikipedia. Our focus is to enable machine to understand technical materials those are contained in scientific articles. Reading technical and scientific texts involves a complex process. It is entirely different from reading general content. So, MRC on scholarly articles is more challenging compared to the other domains. Human being generally put their level best intelligence when they write a research article. It is very difficult to decode the content of the articles even for the human being when they do review or read for research purpose.

Readers, while reading a research article, generally read the abstract and conclusion before delving into the details of the article. This is also true for the editors or associate editors who go through the abstract of the article to get an idea about the quality of the paper as well as assigning appropriate reviewers. In this paper we first create a dataset for MRC in research article, and then develop a deep learning based model as a baseline for further research. The underlying principle is to develop an AI Assisted Peer Review System which will help the editors and reviewers by providing the answers to some basic questions. The datasets for this kind of tasks typically contain *document-question-answer* triples. The answer of the question can take very different forms, depending upon the answer types. Typically, existing MRC tasks can be divided into four categories (Chen, 2018) depending upon the answer type: *i. Cloze Style, ii. Multiple Choice, iii. Span Prediction, and iv. Free-Form Answer*. In general, this research has very recently attracted the attention of the researchers. In particular, we find only two prior works towards this direction (i.e. on research articles). The first one (Kim et al., 2018) made use of biomedical journals and the second one (Park et al., 2019) fostered scientific journals. But both of these works are based on Cloze Style MRC task. In contrast, our task focuses on preparing a dataset for span-of-words prediction (Span Prediction) based MRC model, where the system has to extract span-of-words as answer to a particular question based on the context. We employ articles from Elsevier Computer Science Journals (like ARTINT, COMNET etc.).

1.1. Related Work

The problem of document understanding falls in the domain of Natural Language Understanding (NLU), and has a long history. Machine Reading Comprehension (Hermann et al., 2015) and Open Domain Question Answering (Chen et al., 2017a) are the two very challenging tasks and fall under the domain of NLU. To encourage this task, over the years research community has come up with publicly available several datasets and methods for benchmarking. We condense a few significant of them, and show in Table 1. We describe these briefly in the following:

The Stanford Question Answering Dataset (SQuAD): Rajpurkar et al. (2016) presented the RC dataset having more than 100k questions constructed by the crowdworkers on a set of Wikipedia articles. The second version of SQuAD was released by Rajpurkar et al. (2018) that focuses on unanswerable questions. This version combines the previous version of SQuAD and additionally over 50,000 unanswerable questions are written adversarially by crowdworkers to look into the similar ones.

MAchine Reading COmprehension Dataset (MS-MARCO): Nguyen et al. (2016) proposed a dataset that comprises of 1 million anonymized questions sampled from Bing's search query logs.

NewsQA: This dataset (Trischler et al., 2017) consists of more than 100,000 human generated QA pairs. The goal of preparing this was to test the MRC models on reasoning skills.

DuReader: This is a Chinese MRC dataset proposed by He et al. (2018). The dataset was created with the real application data from Baidu search and Baidu Zhidao (a community QA website). It comprises of 200,000 questions and 420,000 answers from 1,000,000 documents. In this dataset the answers have additional label like either fact based or opinionative.

NarrativeQA: Kočiskỳ et al. (2018) created NarrativeQA based on summaries of movie scripts and books. Previous datasets and models are controled by questions that can be answered by selecting answers using local contextual similarity or global term frequency. This dataset encourages the deeper understanding of languages.

Span Extract Chinese MRC Dataset: Cui et al. (2019) recently proposed a novel dataset to add language diversities in this area as the existing datasets focus on only English language. The dataset is composed by near 20,000 real questions annotated on Wikipedia paragraphs by human experts.

Delta Reading Comprehension Dataset (DRCD): Shao et al. (2018) proposed an open domain traditional Chinese MRC dataset. The main aim of this dataset is to be a standard Chinese MRC dataset, which could be utilised as a source dataset for transfer learning. It comprises of 10,014 paragraphs obtained from 2,108 Wikipedia articles and from there 30,000+ questions generated by annotators.

RACE: This dataset was created by Lai et al. (2017), and contains nearly 100,000 multiple choice questions and 27,000 passages from standardized tests for Chinese students, who are learning English as a foreign language. The aim of creating this dataset is to test the students′ ability in understanding and reasoning, covering variety of topics.

AI2 Reasoning Challenge (ARC): A team (Clark and Gardner, 2018) of Allen Institute for Artificial Intelligence prepared this dataset. It consists of 7,787 grade-school multiple choice (with 4 possible options) science question.

ReCoRD: Zhang et al. (2018) represents this MRC dataset that requires commonsense reasoning. It contains 12,000 cloze-style question passage pairs extracted from CNN/Daily Mail news articles. It requires common sense reasoning to answer the queries of this dataset.

1.2. Motivation and Contribution

Our understanding and survey reveal that- although there are many benchmark datasets available for question-answering- but there has not been any significant effort for building models related to the domain of research articles. The ultimate goal is to build an AI Assisted Peer Review System. This would provide assistance to the editors and reviewers by providing answers to the basic questions related to the research article. The questions could be:

- *What problem does the article attempt to address?*
- *What is the method used?*
- *What was the underlying motivation?*
- *Is the evaluation done on benchmark dataset?*
- *Are the results state-of-the-art?*
- *Are the results significant?*

Based on the answers to these questions, editor will get an idea about the research article, and this will enable them to make an appropriate decision, i.e. either desk-rejecting it or forwarding to the next level and assigning the appropriate reviewers. Our contributions could be outlined as follows:

- We create a benchmark dataset for scholarly article reading comprehension. This is span-of-words-based reading comprehension dataset. To the best of our knowledge, this is the very first attempt towards this direction.
- We set a baseline model by building a deep learning based machine reading comprehension system based on the BiDAF framework.

Rest of the paper is arranged as follows. We first discuss the data creation process in Section 2. Section 2 describes the evaluation of generated data and data annotation guidelines. We present the neural baseline model in Section 3. In Section 4, we report the experiments and results. Section 5 concludes the article and some points to future directions of work.

2. Dataset Creation

We consider approximately 300 articles from two Elsevier Computer Science journals, namely Artificial Intelligence (ARTINT), Computer Networks (COMNET). We convert these articles from PDF to XML format using *GeneRation Of Bibliographic Data (GROBID)*[1]. GROBID is a machine learning library for extracting, parsing and re-structuring raw documents such as PDF (specially, technical and scientific publications) into structured XML/TEI encoded documents. Although GROBID is not entirely perfect, however, it performs better compared to the other existing tools, and hence we use this for extracting information from the scientific pdf documents. We extract the abstract of each article from the XML structure. The abstract part of any research article contains abstractive summary of the whole research article. We consider the abstract as the context/document/paragraph for our experiment. We parse the extracted abstracts by the Stanford Constituency Parser[2]. We take the implementation available in Stanford CoreNLP for Constituency Parser. Please note that the parser is having a test F1 of 91.3% (Zhu et al., 2013). The Constituency Parser essentially breaks each sentence of the abstract into noun and verb phrases. We extract the noun phrases for a particular abstract. We consider these noun phrases as the plausible answers for that particular context/document. It has been observed from the literature (Rajpurkar et al., 2016; Trischler et al., 2017) that most of the answers of a particular document lies in the category of noun phrase.

The study of Rajpurkar et al. (2016) showed diversity in answer by categorizing them. They categorise the answers numerically and non-numerically. Non-numerical answers are

[1] https://grobid.readthedocs.io/en/latest/Introduction/

[2] https://nlp.stanford.edu/software/lex-parser.shtml

Dataset	Question Source	Answer	Size	Domain
ScholarlyRead (Proposed)	Semi-Automatic	Span of Words	10K	Scholarly Articles
BioRead (Pappas et al., 2018)	Cloze	Fill in single word	16.4 million	Bio-Medical Literature
SQuAD (Rajpurkar et al., 2016)	Crowd-sourced	Span of words	100K	Wikipedia
TREC-QA (Voorhees and Tice, 2000)	Query Logs	IR, Free Form	1479	Short answer questions from any domain
WikiQA (Yang et al., 2015)	Bing Query Logs	IR, Sentence selection	3047	Wikipedia
Algebra (Kushman et al., 2014)	Standardized tests	Computation	514	Algebra word problems
Science (Clark and Etzioni, 2016)	Standardized tests	multiple choice	855	Math. and Science Test
NewsQA (Trischler et al., 2017)	Crowd-sourced	Span of Words	100k	News
DuReader (He et al., 2018)	Crowd-sourced	Human Generated	200K	Chinese Document.
Narrative QA (Kočiský et al., 2018)	Crowd-sourced	Human Generated	46,765	books and movie scripts
MC Test (Richardson et al., 2013)	Crowd-sourced	Multiple choice	2640	Fictional story
CNN/Daily Mail (Chen et al., 2016)	Cloze+Summary	Fill in single word	1.4M	News Articles
CBT(Hill et al., 2015)	Cloze	Fill in single word	688k	freely available cultural eBooks

Table 1: A Comparison of existing MRC and QA Datasets. ScholarlyRead is different from others in terms of domain.

categorized using constituency parses and POS tags generated by the Stanford CoreNLP. The proper noun phrases are further split into person, location, and other entities using the Named Entity (NE) information. They made the following analysis of the answers: 32.6% proper noun, 31.8% common noun, 19.8% are dates and other numbers, 15.8% of the answers are adjective phrase. Another study (Trischler et al., 2017) worked following the same line. This study followed the same procedure as the previous one and made analysis of the answers as follows: 22.2% are common noun phrase, 18.3% clause phrase, 14.8% person, 9.8% numeric, 11.2% other types. Taking the concepts from these two articles we consider the extracted noun phrases as the answers for a particular document. So, now we have context (abstract) and its plausible answers. We pair the context and it's answers. The context answer pairs are given to a pre-trained question generation model. We use the Stanford Question Answering Dataset (SQuAD) [3] for the training purpose. We develop the model based on Yuan et al. (2017). Hence, the model takes context/document and answer pairs as input, and as output it produces question for that particular answer from the document. The model makes use of simple encoder-decoder model as outlined in Cho et al. (2014). It makes use of the combination of supervised and reinforcement learning for the training purpose. Taking the output from this model we are having the triples of *Document-Answer-Question*. These generated questions are manually checked by the annotators. In this way we create the dataset for Reading Comprehension (RC) on scholarly articles and coin it as ***ScholarlyRead*** [4].

We depict the flowchart of the question generation model in Figure 1.

2.1. Evaluation of Generated Questions

We employ two annotators for the evaluation of the generated questions. They were asked to create reference questions manually given the document and answer. We compare the system generated questions with the reference questions. We took a sample of 1500 instances and performed evaluation. The results are shown in Table 2. BLEU, METEOR and ROUGE are the three major and widely used Natural Language Generation (NLG) evaluation schemes. We also use different versions of BLEU (like BLEU_2, BLEU_3 etc) to obtain bi-grams, tri-grams matching in addition to uni-gram matching. Apart from these we also use the logic based evaluation. We use entailment model for this purpose. We make use of Enhanced Sequential Inference model (Chen et al., 2017b), which is one of the state-of-the-art and widely used entailment models. The model comprises of three modules: Input Encoding, Local Inference Modeling, and Inference Composition. It demonstrates that carefully designing sequential inference model based on chain LSTM can perform better than the other models. The model is trained on the Stanford Natural Language Inference (SNLI) dataset [5]. Annotators were asked to provide entailment labels (Entailment, Contradiction and Neutral) to each reference question with respect to the system generated question. We run this model between system and human generated questions. System predicts the entailment label to each such instances. We compare with the gold labels as given by the human. We obtain an accuracy of 56%. The score indicates that the generated questions are logically entailed with the human generated questions.

BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	ROUGE_L
0.29	0.18	0.13	0.10	0.17	0.26

Table 2: Evaluation results on different unsupervised automated metrics for NLG

2.2. Annotation Guidelines

In order to check the quality of generated questions, we employ two annotators. The annotators are postgraduate in language, and have good expertise in the related field. While checking the system generated questions they were given the following instructions:

- The questions should be grammatically correct, which include:

[3] https://rajpurkar.github.io/SQuAD-explorer/

[4] https://www.iitp.ac.in/~ai-nlp-ml/resources.html#ScholarlyRead

[5] https://nlp.stanford.edu/projects/snli/

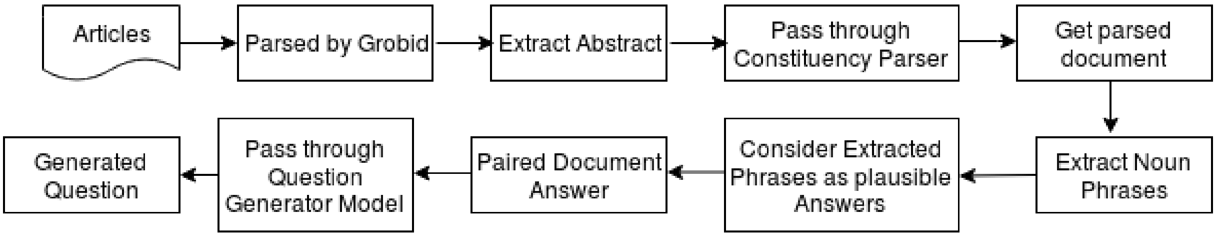

Figure 1: Automatic Question Generation Approach

- No conjunction
- Spelling
- Punctuation
- Upper casing of proper names and beginning of the sentences

- The questions should be fully interpretable/understandable even for the reader who doesn't know the context of the corresponding passage.

- As all of our questions are factoid in nature, they were also asked to check whether the answers to the questions are unique and factual or not.

We randomly picked up 100 samples. The annotators were asked to assign the score between 1-5 based on the Naturalness, which includes Grammaticality and Fluency. The inter-annotator agreement ratio was found to be 0.71 in terms of Kappa Coefficient (Cohen, 1960). This may be considered to be a substantial agreement according to Landis and Koch (1977).

3. Method

The dataset created (i.e., ScholarlyRead dataset) could be served as the dataset for building Question Answering (QA) models on scholarly articles. In particular, it could be a benchmark dataset for span-of-words-based MRC systems in the domain of scholarly articles. The features of the MRC model could be as follows:

- The system would be able to learn to comprehend Scholarly articles. After comprehending, it will be able to give answers to the questions based on the article.

- The answers to questions are essentially the phrases (span-of-words) from passage. Hence, the model has to predict the start and end index of the phrase.

- The kind of MRC model that we propose here would help the editor of any journal or chair/co-chairs of any conference to expedite the automation process of article classification and/or automatic article review systems.

- Additionally, this MRC model would assist the editor in getting the answers of some basic questions as mentioned in Section 1.2.

- The system will enable the editor to get a fine-grained view of an incoming article.

	# of Examples		
Article	Train	Dev	Test
	8500	1500	500

Table 3: Statistics of the used dataset for training and testing of the BiDAF method

- Accordingly, the editor will be able to choose the reviewers for the reviewing purposes.

- As per general practice, the editors classify the articles based on the authors' given keywords. Sometimes, authors may give wrong keywords (as the people do not give the keywords so seriously). The editor send the articles accordingly to the domain experts. Now this kind of system would give the editors a more clear picture about the incoming articles.

- Similarly, the researchers could be able to get answers to those elementary questions before reading and/or going into detail a particular article.

We implement a method that is based on the Bi-directional Attention Flow (BiDAF) (Seo et al., 2016) network. It is one of the simplest and widely used well-known deep learning based MRC models for span-of-words-based QA system. It is observed that increasing training set of BiDAF leads to much larger performance gain. We consider pytorch implementation of this model [6].
BiDAF: It takes Context and Question as input, and predicts the start and end indices of a phrase, which is considered to be the predicted answer of the question. It includes character, word, and contextual level embedding. It also uses bi-directional attention flow to obtain a query-aware context representation. The proposed attention mechanism provides several advantages to the previously proposed attention mechanisms. The attention is computed for every time step, and the attended vector at each time step, along with the representations from the previous layers, is allowed to flow through to the subsequent modeling layer. Secondly, this model used memory-less attention mechanism. It uses dual attention i.e. one attention from query to context and the other from context to query, that provides complementary information to each other.

4. Experiments and Results

We run the BiDAF model on our dataset. Statistics of training and testing sets are shown in Table 3.
We make use of two standard separate metrics to evaluate

[6] https://github.com/galsang/BiDAF-pytorch

Metric	%
Exact Match	20.6
Micro-averaged F1 score	37.31

Table 4: Results obtained on scholarly article dataset

the model's performance as used in (Rajpurkar et al., 2016). The metrics ignore punctuation and articles (e.g. a, an, the etc). We are having only one reference answer for testing.

Exact Match: We measure the percentage of predictions that match exactly with the ground truth answer.

F1 score (Macro-averaged): We treat the predicted and ground truth answers as bags of tokens. We then measure the average overlap between the predicted and ground truth answer, compute their F1, and then take the average over all of the examples.

We show the results using these evaluation metrics in Table 4. This paper focuses on scientific texts. Research articles are the manifestation of highest form of human intelligence. The computerised processing and comprehending such intelligent texts are very complex in nature, compared to other domains. It is hard task, even for a human being. Low scores obtained by the model indicate this, and opens the door for further research in this domain to the community. The word embedding method (Glove (Pennington et al., 2014)) used in BiDAF model is trained on other domain's vocabulary. So it is obvious that representation obtained for a particular scientific text would not be accurate. This could be one of the reasons for the poor performance of our baseline model's.

5. Conclusion and Future Works

In this paper we have presented a benchmark setup for reading comprehension on scholarly articles. We create the dataset, ScholarlyRead, a novel span-of-words-based reading comprehension dataset on scholarly articles. We have created this dataset from the various computer science related journals of Elsevier (like ARTINT, COMNET etc). We follow a semi-automatic way to generate the questions. Firstly, we generate the questions from a pre-trained model, and then perform a manual verification. We have developed a span-of-words-based RC model, that is based on BiDAF. We set this model as a baseline for future research. A deeper analysis shows that more investigations are needed to improve the quality. The future work includes (i). considering all the Elsevier Computer Science Journals for the enrichment of this dataset; (ii). incorporating multiple hop attention in BiDAF model; (iii). developing multi-hop version of this dataset; (iv). considering the full text of a research article instead of only abstract; and (v). as the dataset created on closed domain, as an extension of this work, we propose to train the model on the already generated data and also start working on the open access papers. The test cases could be the open-access data (say from AI journal, AAAI open access papers). In this way we could preserve the domain dependency.

6. Acknowledgements

We would like to acknowledge ”Elsevier Centre of Excellence for Natural Language Processing” at Computer Science Department in Indian Institute of Technology Patna for supporting the research work furnished over here in this article. We also acknowledge Mr. Pallab Bhattacharjee for helping us in manual checking of the generated questions, and Mr. Deepak Gupta (PhD Scholar) for helping in implementation.

7. Bibliographical References

Chen, D., Bolton, J., and Manning, C. D. (2016). 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, Berlin, Germany, August. Association for Computational Linguistics.

Chen, D., Fisch, A., Weston, J., and Bordes, A. (2017a). Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada, July. Association for Computational Linguistics.

Chen, Q., Zhu, X., Ling, Z.-H., Wei, S., Jiang, H., and Inkpen, D. (2017b). Enhanced LSTM for Natural Language Inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada, July. Association for Computational Linguistics.

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

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October. Association for Computational Linguistics.

Clark, P. and Etzioni, O. (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.

Clark, C. and Gardner, M. (2018). Simple and Effective Multi-Paragraph Reading Comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 845–855, Melbourne, Australia, July. Association for Computational Linguistics.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and psychological measurement*, 20(1):37–46.

Cui, Y., Liu, T., Che, W., Xiao, L., Chen, Z., Ma, W., Wang, S., and Hu, G. (2019). A Span-Extraction Dataset for Chinese Machine Reading Comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5886–5891, Hong Kong, China, November. Association for Computational Linguistics.

He, W., Liu, K., Liu, J., Lyu, Y., Zhao, S., Xiao, X., Liu, Y., Wang, Y., Wu, H., She, Q., Liu, X., Wu, T., and Wang, H. (2018). DuReader: A Chinese Machine Reading Comprehension Dataset from Real-world Applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46, Melbourne, Australia, July. Association for Computational Linguistics.

Hermann, K. M., Kociský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching Machines to Read and Comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

Hill, F., Bordes, A., Chopra, S., and Weston, J. (2015). The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. *arXiv preprint arXiv:1511.02301*.

Kim, S., Park, D., Choi, Y., Lee, K., Kim, B., Jeon, M., Kim, J., Tan, A. C., and Kang, J. (2018). A Pilot Study of Biomedical Text Comprehension using an Attention-Based Deep Neural Reader: Design and Experimental Analysis. *JMIR medical informatics*, 6(1):e2.

Kočiskỳ, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K. M., Melis, G., and Grefenstette, E. (2018). The NarrativeQA Reading Comprehension Challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Kushman, N., Artzi, Y., Zettlemoyer, L., and Barzilay, R. (2014). Learning to Automatically Solve Algebra Word Problems. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 271–281.

Lai, G., Xie, Q., Liu, H., Yang, Y., and Hovy, E. (2017). RACE: Large-scale ReAding Comprehension Dataset from Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark, September. Association for Computational Linguistics.

Landis, J. R. and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *biometrics*, pages 159–174.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. (2016). MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*.

Pappas, D., Androutsopoulos, I., and Papageorgiou, H. (2018). BioRead: A New Dataset for Biomedical Reading Comprehension. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Park, D., Choi, Y., Kim, D., Yu, M., Kim, S., and Kang, J. (2019). Can Machines Learn to Comprehend Scientific Literature? *IEEE Access*, 7:16246–16256.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). 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, Austin, Texas, November. Association for Computational Linguistics.

Rajpurkar, P., Jia, R., and Liang, P. (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia, July. Association for Computational Linguistics.

Richardson, M., Burges, C. J., and Renshaw, E. (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, Seattle, Washington, USA, October. Association for Computational Linguistics.

Seo, M., Kembhavi, A., Farhadi, A., and Hajishirzi, H. (2016). Bi-Directional Attention Flow for Machine Comprehension. *arXiv preprint arXiv:1611.01603*.

Shao, C. C., Liu, T., Lai, Y., Tseng, Y., and Tsai, S. (2018). DRCD: A Chinese Machine Reading Comprehension Dataset. *arXiv preprint arXiv:1806.00920*.

Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., and Suleman, K. (2017). NewsQA: A Machine Comprehension Dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada, August. Association for Computational Linguistics.

Voorhees, E. M. and Tice, D. M. (2000). Building a Question Answering Test Collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 200–207, New York, NY, USA. ACM.

Yang, Y., Yih, W. t., and Meek, C. (2015). WikiQA: A Challenge Dataset for Open-Domain Question Answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal, September. Association for Computational Linguistics.

Yuan, X., Wang, T., Gülçehre, Ç., Sordoni, A., Bachman, P., Zhang, S., Subramanian, S., and Trischler, A. (2017). Machine Comprehension by Text-to-Text Neural Question Generation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017*, pages 15–25.

Zhang, S., Liu, X., Liu, J., Gao, J., Duh, K., and Van Durme, B. (2018). ReCoRD: Bridging the Gap between Human and Machine Commonsense Reading Comprehension. *arXiv preprint arXiv:1810.12885*.

Zhu, M., Zhang, Y., Chen, W., Zhang, M., and Zhu, J. (2013). Fast and Accurate Shift-Reduce Constituent Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 434–443, Sofia, Bulgaria, August. Association for Computational Linguistics.