Frustratingly Hard Evidence Retrieval for QA Over Books

Xiangyang Mou
Rensselaer Polytechnic Institute
Troy, NY 12180
moux4@rpi.edu

Mo Yu
IBM Research
USA
yum@us.ibm.com

Bingsheng Yao
Rensselaer Polytechnic Institute
Troy, NY 12180
yaob@rpi.edu

Chenghao Yang
Columbia University
New York, NY 10027
chenghao.yang@columbia.edu

Xiaoxiao Guo
IBM Research
USA
xiaoxiao.guo@ibm.com

Saloni Potdar
IBM Watson
USA
potdars@us.ibm.com

Hui Su
IBM Research
USA
huisuibmres@us.ibm.com

Abstract

A lot of progress has been made to improve question answering (QA) in recent years, but the special problem of QA over narrative book stories has not been explored in-depth. We formulate BookQA as an open-domain QA task given its similar dependency on evidence retrieval. We further investigate how state-of-the-art open-domain QA approaches can help BookQA. Besides achieving state-of-the-art on the NarrativeQA benchmark, our study also reveals the difficulty of evidence retrieval in books with a wealth of experiments and analysis - which necessitates future effort on novel solutions for evidence retrieval in BookQA.

1 Introduction

The task of question answering has benefited largely from the advancements in deep learning, especially from the pre-trained language models(LM) (Radford et al., 2019; Devlin et al., 2019). While question answering over single passage (reading comprehension datasets) and over the large-scale open-domain corpora (open-domain QA) have largely benefited from these, the performance of QA over book stories (BookQA) lags behind. For example, the most representative benchmark in this direction, the NarrativeQA (Kočiskỳ et al., 2018) which was released three years ago - the current state-of-the-art methods only show marginal improvement over the first baselines.

There are several challenges in NarrativeQA which slow down the research progress. First, the narrative stories lead to a new writing style which differs from previous works over formal texts like Wikipedia. Second, the long inputs of books are beyond the processing ability of neural models so evidence identification from a whole book is critical. Third, NarrativeQA is a generative task, and many of the answers cannot be exactly matched in the original books. Hence, the generative QA models are required. Finally and most importantly, the dataset does not provide annotations of the supporting evidence. While this makes it a realistic setting like open-domain QA, together with the generative nature of the answers, also makes it difficult to infer the supporting evidence similar to most of the extractive open-domain QA tasks.

The requirements around evidence identification and the missing supporting evidence annotation make BookQA task similar to open-domain QA. In this paper, we first study whether the ideas used in state-of-the-art open-domain QA systems can be extended to improve BookQA including: (1) the neural ranker-reader pipeline (Wang et al., 2018), where a neural ranker is used to select related passages (evidence) given a question from a large candidate sets; (2) the usage of pre-trained LMs as reader and ranker, such as GPT (Radford et al., 2019), BERT (Devlin et al., 2019) and their follow-up work; (3) the distantly supervised and unsupervised training techniques (Wang et al., 2018; Lee et al., 2019; Min et al., 2019; Guu et al., 2020; Karpukhin et al., 2020) that help rankers learn more from noisy gold data.

By training a ranker-reader framework on BookQA, we successfully achieve a new state-of-the-art on NarrativeQA using both generative and extractive readers. Based on these results and our

analysis, we observe the followings:

• Using the pre-trained LMs as the reader model, such as BERT and GPT, improves the NarrativeQA performance. With the same BM25 IR baseline, they give 5-6% improvement on Rouge-L over their non-pre-trained counterparts.

• Our specifically designed distant supervision signals improve the neural ranker significantly, but the improvement is small compared to the upper bound. Further analysis of the ranker module confirms the difficulty in training, as the improvement from the pre-trained LM BERT is marginal in it.

2 Proposed Method

2.1 Task Definition

Following (Kočiskỳ et al., 2018), we define the task of **BookQA** as finding the answer $\mathbf{A}$ to a question $\mathbf{Q}$ from a book $\mathbf{B}$,[1] where each book contains a number of consecutive paragraphs $\mathcal{C}$ (usually hundreds or more). $\mathbf{A}$ is a free-form answer that can be concluded from the book but may not appear in it in an exact form.

In this paper we propose an open-domain QA formulation and solution to the task of BookQA. Specifically, the task consists of (1) an evidence retrieval step that selects evidence from $\mathbf{B}$ for $\mathbf{Q}$, which in our case is a collection of paragraphs $\mathcal{C}_\mathbf{Q} = \{\mathbf{C_i}\} \subset \mathbf{B}$; and (2) a question-answering step that predicts an answer given $\mathbf{Q}$ and $\mathcal{C}_\mathbf{Q}$.

In the state-of-the-art open-domain QA systems, the aforementioned two steps are modeled by two learnable models (usually based on pre-trained LMs), namely the **ranker** and the **reader**. The ranker predicts the relevance of each paragraph $\mathbf{C} \in \mathbf{B}$ to the question, where the top ranked paragraphs form the $\mathcal{C}_\mathbf{Q}$; and the reader predicts the answer following $P(\mathbf{A}|\mathbf{Q}, \mathcal{C}_\mathbf{Q})$.

In the following subsections, we describe our solution to make the training of pre-trained LM-based ranker and reader work for the BookQA task.

2.2 Reader (QA Model)

Extractive Reader We use a pre-trained BERT model (Devlin et al., 2019; Wolf et al., 2019) to predict the answer span given the query and the context. One challenge of training an extraction model in BookQA is that there is no annotation of true spans because of its generative nature. Our solution is to find the most likely span as answer supervision. Specifically, we compute the Rouge-L score (Lin, 2004) between the true answer and each candidate span of the same length, and finally take the span with the maximum Rouge-L score as our weak label. We initially tried the exact-answer spans but failed to find many due to its low coverage in BookQA.

Generative Reader Considering the GPT memory limitation, we use the `GPT-2-medium` model as our pre-trained generative model and fine-tune it on BookQA using default training parameters[2].

2.3 Book Paragraph Ranker

We fine-tune another BERT binary classifier for paragraph retrieval, following the usage of BERT on text similarity tasks. In BookQA, training such a classifier is challenging because of the lack of evidence-level supervision. We deal with this problem by using an ensemble method to achieve distant supervision. We build two weak BM25 retrievers with one using only $\mathbf{Q}$ and the other using both $\mathbf{Q}$ and true $\mathbf{A}$. Denoting the correspondent rough-grained retrievals as $\mathcal{C}_\mathbf{Q}$ and $\mathcal{C}_{\mathbf{Q}+\mathbf{A}}$, we then tutor a model to select their intersection $\mathcal{C}_\mathbf{Q} \cap \mathcal{C}_{\mathbf{Q}+\mathbf{A}}$ by sampling the positive samples from $\mathcal{C}_\mathbf{Q} \cap \mathcal{C}_{\mathbf{Q}+\mathbf{A}}$ and the negative ones from $(\mathcal{C}_\mathbf{Q} \cap \mathcal{C}_{\mathbf{Q}+\mathbf{A}})^\mathsf{c}$. In order to encourage the ranker to select passages that have better coverage of the answers, we further apply a **Rouge-L filter** upon the previous sampling results, and only select the positive samples whose answer-related Rouge-L score is higher than the upper threshold and the negative samples lower than the lower threshold[3].

3 Experiments

3.1 Settings

Dataset We conduct experiments on NarrativeQA dataset (Kočiskỳ et al., 2018), which has a collection of 783 books and 789 movie scripts and their summaries, with each having on average 30 question-answer pairs. Each book or movie script contains an average of 62k words. NarrativeQA provides two different settings, the **summary** setting and the **full-story** setting. Our BookQA task corresponds to the full-story setting that finds answers from books or movie scripts. Note that the NarrativeQA is a *generative* QA task. The answers are not guaranteed to appear in the books.

[1] To be more accurate, the question should be denoted as $\mathbf{Q_B}$ but we use $\mathbf{Q}$ for simplicity.

[2] https://huggingface.co/transformers/model_doc/gpt2.html

[3] In practice, we set the hyperparameters 0.7 and 0.4

System	w/ trained ranker	w/ pre-trained LM	w/ extra training data
Attention Sum (Kočiskỳ et al., 2018)			
BiDAF (Kočiskỳ et al., 2018)			
IAL-CPG (Tay et al., 2019)			
R^3 (Wang et al., 2017)	✓		
BERT-heur (Frermann, 2019)	✓	✓	✓
Our generative/extractive systems	✓	✓	

Table 1: Summary of the characteristics of the compared systems. Red/blue color refers to generative/extraction QA systems. In addition to the standard techniques, (Wang et al., 2017) uses reinforcement learning to train the ranker; and (Tay et al., 2019) uses curriculum to train the reader to overcome the divergence of evidence retrieval qualities between training and testing.

We preprocess the raw data with SpaCy[4] tokenization. Then following (Kočiskỳ et al., 2018), we cut the books into non-overlapping paragraphs with a length of 200 each for the full-story setting.

Baseline We conduct experiments with both generative and extractive readers, and compare with the competitive baseline models from (Kočiskỳ et al., 2018; Tay et al., 2019; Frermann, 2019) in the full-story setting. Meanwhile, we take a BM25 retrieval as the baseline ranker and evaluate our distantly supervised BERT rankers. We also compare to the strong results from (Frermann, 2019), which constructed evidence-level supervision with the usage of book summaries. However, the summary is not considered available by design (Kočiskỳ et al., 2018) in the general full-story scenario where questions should be answered solely from books.[5]

Although not the focus of the paper, our reader performance in the summary setting is also reported (Section 3.2), to show the properties of the readers.

Metrics Because of the generative nature of the task, following previous works (Kočiskỳ et al., 2018; Tay et al., 2019; Frermann, 2019), we evaluate the QA performance with Bleu-1, Bleu-4 (Papineni et al., 2002), Meteor (Banerjee and Lavie, 2005), Rouge-L (Lin, 2004).[6] We also report the Exact Match(EM) and F1 scores[7] that are commonly used in open-domain QA evaluation. We convert both hypothesis and reference to lowercase and remove the punctuation before evaluation.

Model Selection We select the best models on the development set according to its average score of Rouge-L and EM. For ranker model selection, we use the average score of upper bound EM and Rouge-L of top-5 ranked paragraphs.

3.2 Reader Model Validation (the QA-over-Summary Setting)

First, we compare our readers under the summary setting, to verify the correctness of our readers. Our BERT reader achieves performance close to the public state-of-the-art in this setting.

Our GPT-2 reader outperforms the existing systems without usage of pointer generators (PG), but is behind the state-of-the-art with PG. Despite the large gap between systems with and without PG in this setting, (Tay et al., 2019) demonstrates that it didn't contribute much in the full-story setting in the ablation study. Nonetheless, we will investigate the usage of PG in pre-trained LMs in the future work.

3.3 Main Results (the QA-over-Book Setting)

We then experimented our whole QA pipelines in the full-story setting. Table 3 and Table 4 compare our results with public state-of-the-art generative and extractive QA systems.

Our pipeline system with the baseline BM25 ranker outperforms the existing state-of-the-art, confirming the advantage of pre-trained LMs as observed in most QA tasks. Our distantly supervised ranker adds another 1-2% of improvement to all the metrics, bringing both our generative and extractive models with the best performance. It also helps outperform (Frermann, 2019) on multiple metrics without the usage of the strong extra supervision from the summaries.

3.4 Ablation of Ranker Performance

To take a deeper look at the challenges in ranker training, we conduct an ablation study on the ranker independently. The quality of a ranker is measured

[4] https://spacy.io/

[5] In NarrativeQA, the summary has a good coverage of the answers due to the data collection procedures; also, summaries can be viewed as humans' comprehension of the books.

[6] We used an open-source evaluation library (Sharma et al., 2017): https://github.com/Maluuba/nlg-eval.

[7] The squad/evaluate-v1.1.py script is used.

System	Bleu-1	Bleu-4	Meteor	Rouge-L
Extractive Readers				
BERT + Hard EM (Min et al., 2019)	-	-	-	**58.1/58.8**
BERT-only (Min et al., 2019)	-	-	-	55.8/56.1
BERT w/ full training signals [**Ours**]	49.35/49.02	25.76/25.85	23.93/24.14	52.62/52.02
BERT w/ exact answer match only [**Ours**]	**49.78/49.64**	**27.01/28.94**	**25.22/25.12**	57.19/56.35
Generative Readers				
Attention Sum (Kočiskỳ et al., 2018) (w/o PG)	23.54/23.20	5.90/6.39	8.02/7.77	23.28/22.26
Masque (Nishida et al., 2019) (w/ PG)	-/**48.70**	-/**20.98**	-/**21.95**	-/**54.74**
GPT-2 Reader(w/o PG) [**Ours**]	**33.63**/35.49	**11.87**/14.33	**13.71**/14.36	**34.32**/35.65

Table 2: Results under NarrativeQA summary setting on dev/test set (%). PG refers to the usage of pointer generator. For extractive model, we compare with the best public result (Min et al., 2019) and its BERT-only ablation. The latter corresponds to the same setting as ours. For generative model, we compare with the best public models with and without pointer generators.

System	Bleu-1	Bleu-4	Meteor	Rouge-L	EM	F1
Public Generative Baselines						
AttSum (top-10) (Kočiskỳ et al., 2018)	20.00/19.09	2.23/1.81	4.45/4.29	14.47/14.03	-	-
AttSum (top-20) (Kočiskỳ et al., 2018)	19.79/19.06	1.79/2.11	4.60/4.37	14.86/14.02	-	-
IAL-CPG (Tay et al., 2019)	23.31/22.92	2.70/2.47	5.68/5.59	17.33/17.67	-	-
- curriculum	20.75/-	1.52/-	4.65/-	15.42/-		
Our Generative QA Models						
BM25 + GPT-2 Reader	24.54/24.43	4.74/4.37	7.32/7.32	20.25/21.04	5.12/5.22	17.72/18.38
+ BERT Ranker	**24.94/25.03**	**4.76/4.42**	**7.74/7.81**	**21.89/22.36**	**6.79/6.31**	**19.67/19.94**
+ *Oracle IR (BM25 w/ Q+A)*	*33.18/32.95*	*8.16/7.70*	*12.35/12.47*	*34.83/34.96*	*17.09/15.98*	*33.65/33.75*

Table 3: Generative performance in NarrativeQA full-story setting (BookQA setting) dev/test set(%). *Oracle IR* utilizes question and true answers for retrieval.

System	Bleu-1	Bleu-4	Meteor	Rouge-L	EM	F1
Public Extractive Baselines						
BiDAF (Kočiskỳ et al., 2018)	5.82/5.68	0.22/0.25	3.84/3.72	6.33/6.22	-	-
R^3 (Wang et al., 2017)	**16.40/15.70**	0.50/0.49	3.52/3.47	11.40/11.90	-	-
Our Extractive QA Models						
BM25 + BERT Reader	13.27/13.84	0.94/1.07	4.29/4.59	12.59/13.81	4.67/5.26	11.57/12.55
+ BERT Ranker	14.60/14.46	**1.81/1.38**	**5.09/5.03**	**14.76/15.49**	**6.79/6.66**	**13.75/14.45**
+ *Oracle IR (BM25 w/ Q+A)*	*23.81/24.01*	*3.54/4.01*	*9.72/9.83*	*28.33/28.72*	*15.27/15.39*	*28.42/28.55*
Extractive Models w/ additional supervision						
BERT-heur (Frermann, 2019)	-/12.26	-/**2.06**	-/**5.28**	-/15.15	-	-

Table 4: Extractive performance in NarrativeQA full-story setting (BookQA setting) dev/test set(%). *Oracle IR* utilizes question and true answers for retrieval.

by the answer coverage of its top-5 selections on the basis of the top-32 candidates from the baseline. The answer coverage is estimated by the maximum Rouge-L score of the subsequences of the selected paragraphs of the same length as the answers; and whether the answer can be covered by any of the selected paragraphs (EM).

Our BERT ranker together with supervision filtering strategy has a significant improvement over the BM25 baseline. Our BERT ranker improves by 0.7%, compared with MatchLSTM (Wang and Jiang, 2016) or an improved BiDAF architecture (Clark and Gardner, 2018). On the other hand, comparing the benefits that BERT brings to open-domain QA tasks, the relatively small improvement demonstrates the difficulty of evidence retrieval in BookQA. This shows the potential room for future novel improvements, which is also exhibited by the large gap between our best rankers and either the upper bound or the oracle.

3.5 Discussion of Future Improvement

We can see a considerable gap between our best models (ranker and readers) and their correspond-

Question	Gold Answer 1	Gold Answer 2	Generative Result
Where is Millicent sent to boarding school?	Millicent is sent to a boarding school in France	France	France
What is Morgan's relationship to Wyatt?	Morgan is Wyatt's brother	Brothers	Brother
What illness does Doc Holiday suffer from?	Tuberculosis	Tuberculosis	Lung cancer
How does Carl make his house fly?	He attaches thousands of helium balloons to it	Balloons	He uses a parachute to climb up the side of the dirigible
How does Felipe die?	Suicide	He suffers a physical breakdown	He is killed by a bullet in the head
What was the great stone face and how did it appear?	A natural rock formation on the side of a mountain	A natural rock formation which appeared when viewed at a proper distance	It was a stone face

Table 5: Generative result examples. The model tends to generate shorter answers in general. The longer answer it generates, the less likely the answer tends to be correct. The grammatical correctness and fluency of the long generative answers are approaching to human level, regardless of the problematic logic between the generated answer and question. The majority of the generative results do not make sense logically which leads to the low scores in different metrics.

IR Method	EM	Rouge-L
BM25	18.99	47.48
BERT ranker	**24.26**	**52.68**
- Rouge-L filtering	22.63	51.02
Repl BERT w/ BiDAF	21.88	50.64
Repl BERT w/ MatchLSTM	21.97	50.39
Upperbound (BM25 top-32)	30.81	61.40
Oracle (BM25 w/ Q+A)	35.75	63.92

Table 6: IR Evaluation on NarrativeQA dev set(%).

ing oracles in Table 3, 4, and 6. One difficulty that limits the effectiveness of ranker training is the noisy annotation resulted from the nature of the free-form answers. Our filtering technique helps significantly but is still not sufficient. One way we believe that can improve the distant supervision signals is by iteratively updating the ranker and reader like in Hard-EM (Min et al., 2019; Guu et al., 2020). Another possible direction is to extend the idea of inferring evidence on training data with game-theoretic approaches (Perez et al., 2019; Feng et al., 2020), then use the inferred evidence paragraph as labels to train the ranker.

4 Conclusion

We explored the BookQA task and systemically tested on NarrativeQA dataset different types of models and techniques from open-domain QA. Our proposed approaches bring significant improvements to the state-of-the-art across different metrics. Our insight and analysis lay the path for exciting future work in this domain.

Acknowledgment

This work is supported by Cognitive and Immersive Systems Lab (CISL), a collaboration between IBM and RPI, and also a center in IBM's AI Horizons Network.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Christopher Clark and Matt Gardner. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Yufei Feng, Mo Yu, Wenhan Xiong, Xiaoxiao Guo, Junjie Huang, Shiyu Chang, Murray Campbell, Michael Greenspan, and Xiaodan Zhu. 2020. Learning to recover reasoning chains for multi-hop ques-

tion answering via cooperative games. *arXiv preprint arXiv:2004.02393*.

Lea Frermann. 2019. Extractive NarrativeQA with heuristic pre-training. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 172–182, Hong Kong, China. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

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

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. page 10.

Sewon Min, Danqi Chen, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. A discrete hard em approach for weakly supervised question answering. 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 2844–2857.

Kyosuke Nishida, Itsumi Saito, Kosuke Nishida, Kazutoshi Shinoda, Atsushi Otsuka, Hisako Asano, and Junji Tomita. 2019. Multi-style generative reading comprehension. *arXiv preprint arXiv:1901.02262*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Ethan Perez, Siddharth Karamcheti, Rob Fergus, Jason Weston, Douwe Kiela, and Kyunghyun Cho. 2019. Finding generalizable evidence by learning to convince q&a models. In *Proceedings of EMNLP 2019*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799.

Yi Tay, Shuohang Wang, Anh Tuan Luu, Jie Fu, Minh C Phan, Xingdi Yuan, Jinfeng Rao, Siu Cheung Hui, and Aston Zhang. 2019. Simple and effective curriculum pointer-generator networks for reading comprehension over long narratives. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4922–4931.

Shuohang Wang and Jing Jiang. 2016. Learning natural language inference with lstm. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1442–1451.

Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerald Tesauro, Bowen Zhou, and Jing Jiang. 2017. R^3: Reinforced reader-ranker for open-domain question answering.

Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerry Tesauro, Bowen Zhou, and Jing Jiang. 2018. R 3: Reinforced ranker-reader for open-domain question answering. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.