Generative Context Pair Selection for Multi-hop Question Answering

Dheeru Dua[♣] **Cicero Nogueira dos Santos**[♠] **Patrick Ng**[♠]
Ben Athiwaratkun[♠] **Bing Xiang**[♠] **Matt Gardner**[♡] **Sameer Singh**[♣]
[♣]University of California, Irvine, USA
[♠]Amazon Web Services, New York
[♡]Allen Institute for Artificial Intelligence
ddua@uci.edu

Abstract

Compositional reasoning tasks such as multi-hop question answering require models to learn how to make latent decisions using only weak supervision from the final answer. Crowdsourced datasets gathered for these tasks, however, often contain only a slice of the underlying task distribution, which can induce unanticipated biases such as shallow word overlap between the question and context. Recent works have shown that discriminative training results in models that exploit these underlying biases to achieve a better held-out performance, without learning the right way to reason. We propose a generative context selection model for multi-hop QA that reasons about how the given question could have been generated given a context pair and not just independent contexts. We show that on HotpotQA, our proposed generative passage selection model, while being comparable to the state-of-the-art answering performance, has a better performance (4.9% higher than baseline) on an adversarial held-out set that tests the robustness of model's multi-hop reasoning capabilities.

1 Introduction

Recently many reading comprehension datasets like HotpotQA (Yang et al., 2018) and WikiHop (Welbl et al., 2018) that require compositional reasoning over several disjoint passages have been introduced. This style of compositional reasoning, also referred to as multi-hop reasoning, first requires finding the correct set of passages relevant to the question and then finding the answer span in the selected set of passages. These dataset are often collected via crowdsourcing, which makes the training and evaluation of such models heavily reliant on the quality of the collected held-out sets.

Crowdsourced datasets, however, often present only a partial picture of the underlying data distribution. Learning complex latent sequential decisions, like in multi-hop reasoning, to answer a given question under such circumstances is marred by numerous biases, such as annotator bias (Geva et al., 2019), label bias (Dua et al., 2020; Gururangan et al., 2018), survivorship bias (Min et al., 2019b; Jiang and Bansal, 2019), and ascertainment bias (Jia and Liang, 2017). As a result, testing model performance on such biased held-out sets becomes unreliable as the models exploit these biases and learn shortcuts to get the right answer but without learning the right way to reason.

Consider an example from HotpotQA in Figure 1, where the latent entity "Virgina Commonwealth University" can be used by the model (Jiang and Bansal, 2019) to bridge the two relevant passages (highlighted in green) from the original dev set and correctly predict the answer "1838". However, upon adding an adversarial context (highlighted in pink) to the pool of contexts, the model

Question: The 2011-12 VCU Rams men's basketball team, led by third year head coach Shaka Smart, represented the university which was founded in what year?
Gold Answer: 1838

Passage 1: The 2011-12 VCU Rams men's basketball team represented Virginia Commonwealth University during the 2011-12 NCAA Division I men's basketball season...

Passage 2: Virginia Commonwealth University (VCU) is a public research university located in Richmond, Virginia. VCU was founded in 1838 as the medical department of Hampden-Sydney College, becoming the Medical College of Virginia in 1854...

Prediction: 1838

Adversarial context from Jiang and Bansal (2019):
Dartmouth University is a public research university located in Richmond, Virginia. Dartmouth was founded in 1938 as the medical department of Hampden-Sydney College, becoming the Medical College of Virginia in 1854...

New Prediction: 1938

Figure 1: Example from HotpotQA, showing the reasoning chain for answering the question (in green) and an adversarial context (in pink) introduced by Jiang and Bansal (2019) which confuses the model, causing it to change its prediction because it did not learn the correct way to reason.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7009–7015
November 7–11, 2021.

prediction changes to "1938" implying that the model did not learn the right way to reason. This is because the discriminatively trained passage selector exploits lexical cues like "founded" in the second passage and does not pay attention to the complete question. The absence of such adversarial contexts at training allows the model to find incorrect reasoning paths.

In this work, we propose a generative context pair selection model that reasons through the data generation process of how a specific question could have been *generated* given pair of passages. We show that our proposed passage selection module has a better performance on the original (+2.2%) and the adversarial dev set (+4.9%) that tests the model's reasoning abilities (unlike the original dev set which is marred by bias). We use a generic answering model and show that while being comparable in end-to-end performance with close to state-of-the-art systems on the original dev set, our model provides a better performance on the adversarial set. Any advances in the answering model can be applied in a straightforward manner to a generative passage selector to further improve performance.

2 Generative Passage Selection

Given a set of contexts $C = \{c_0, ..., c_N\}$, the goal of *multi-hop* question answering is to combine information from C to an identify answer span a for a given question q. Let $\Psi = \{(c_i, c_j) = c_{ij} : c_i \in C, c_j \in C)\}$ be the set of all possible context pairs that can be formed from C.

Existing models for multi-hop question answering (Tu et al., 2020; Chen et al., 2019) consist of two components: a *discriminative passage selection* and an *answering model*. Passage selection identifies which pairs of contexts are relevant for answering the given question, i.e., it estimates $p(c_{ij} \mid q, \Psi)$. This is followed by the answering model to extract the answer span given a context pair and the question ($p(a \mid q, c_{ij})$). These are combined as follows:

$$p(a \mid q, \Psi) = \sum_{c_{ij}} p(a \mid q, c_{ij}) p(c_{ij} \mid q, \Psi) \quad (1)$$

The discriminative passage selector learns to select a set of contexts by conditioning on the question representation. This learning process does not encourage the model to pay attention to the entire question, which can result in ignoring parts of the question, and thus, learning spurious correlations.

For prediction, best context pair c^*_{ij} is used by the answering module to get the answer, $a^* = \text{argmax}\, p(a \mid q, c^*_{ij})$. As shown by Min et al. (2019a), using the top scoring reasoning chain to answer the question is often sufficient and does not require marginalization over multiple chains.[1]

2.1 Proposed Model

We propose a joint question-answering model that learns $p(a, q \mid \Psi)$ instead of $p(a \mid q, \Psi)$. This model can be factorized into a generative passage selector and a standard answering model as:

$$p(a, q \mid \Psi) = \sum_{c_{ij}} p(a \mid q, c_{ij}) p(q|c_{ij}) p(c_{ij}|\Psi) \quad (2)$$

A prior $p(c_{ij}|\Psi)$ over the context pairs establishes a measure of compatibility between passages in a particular dataset. The conditional generation model $p(q|c_{ij})$ estimates the likelihood of generating the given question from a selected pair of passages. Finally, a standard answering model $p(a \mid q, c_{ij})$ learns a likely answer distribution given a question and context pair. The first two terms (prior and conditional generation) can be seen as a generative model that selects a pair of passages from which the question could have been constructed. The answering model can be instantiated with any existing SOTA model, like graph neural network (Tu et al., 2020; Shao et al., 2020) and entity-based chain reasoning (Chen et al., 2019).

The process at prediction is identical to that with discriminative passage selection, except that the context pairs are scored by taking the entire question into account, $c^*_{ij} = \text{argmax}_{c_{ij}}\, p(q|c_{ij}) p(c_{ij}|\Psi)$.

2.2 Model Learning

For learning the generative model, we train the prior, $p(c_{ij}|\Psi)$ and the conditional generation model $p(q \mid c_{ij}, \Psi)$ jointly. First, the prior network projects the concatenated contextualized representation, r_{ij}, of starting and ending token of concatenated contexts $(c_i; c_j)$, from the encoder to obtain un-normalized scores, which are then normalized across all context-pairs via softmax operator. The loss function tries to increases the likelihood of gold context pair over all possible context pairs.

$$r_{ij} = encoder(c_i; c_j) \quad (3)$$

$$s_{ij} = W^{1\times d}(r_{ij}[start]; r_{ij}[end]) \quad (4)$$

[1] Marginalizing over all context pairs, or maintaining a beam of highly ranked pairs, did not yield much higher performance, in particular, not worth the computation overhead.

The conditional question generation network gets contextual representations for context-pair candidates from the encoder and uses them to generate the question, via the decoder. The objective function increases the likelihood of the question for gold context pairs and the unlikelihood (Welleck et al., 2020) for a set of *negative* context pairs (Eq. 5). The negative context pairs are randomly sampled from all possible non-oracle context pairs.

$$\mathcal{L}(\theta) = \sum_{t=1}^{|question|} \log p(q_t \mid q_{<t}, c_{gold}) + \sum_{n \in |neg.pairs|} \sum_{t=1}^{|question|} \log\left(1 - p(q_t \mid q_{<t}, c_n)\right) \quad (5)$$

3 Experiments and Results

We experiment with two popular multi-hop datasets: HotpotQA (Yang et al., 2018) and WikiHop (Welbl et al., 2018). We use a pre-trained T5 (Raffel et al., 2019) encoder-decoder model for obtaining contextual representations, which are further trained to estimate all individual probability distributions. The answering model is a fine-tuned T5-large model which has an *oracle* EM/F1, $p(a \mid q, c_{gold})$, of 74.5/83.5 and 76.2/83.9 on HotpotQA and WikiHop respectively. The performance of a fine-tuned T5-base (220M parameters) model for standard and generative passage selector are shown in Table 2. Most SOTA passage selectors for HotpotQA use a RoBERTa-large (Liu et al., 2019) (355M parameter) based classifier to select top-k passages given the question, which has an accuracy of ~94.5% (Tu et al., 2020).

Table 1 compares end-to-end original dev set performance of answering model when combined with a standard (79.5 F_1) and a generative passage selector (81.9 F_1) for HotpotQA. Table 2 shows minor improvements in passage accuracy on using generative selector in WikiHop. This shows that generative passage selecetor is able to find (latent) entity connections between context pairs that are consistent with the complete question and not just parts of it.

3.1 Adversarial Evaluation

We use an existing adversarial set (Jiang and Bansal, 2019) for HotpotQA to test the robustness of model's multi-hop reasoning capabilities given a confusing passage. This helps measure, quantitatively, the degree of biased correlations learned by the model. In Table 1, we show that the standard discriminative passage selector has a much higher performance drop (~4%) as compared to the generative selector (~1%) on adversarial dev set (Jiang and Bansal, 2019), showing that generative selector is less biased and less affected by conservative changes (Ben-David et al., 2010) to the data distribution. While the end to end QA performance of our model is comparable (↓ 0.3 F1) to Fang et al. (2020) on the original dev-set, on the adversarial set our method is better than Fang et al. (2020) (↑ 1.2 F1). Table 3 shows that the decoder of generative passage selector was able to generate multi-hop style questions from a pair of contexts.

Model	Original		Adversarial	
	Acc	F_1	Acc	F_1
Standard Selector	95.3	79.5	91.4	76.0
Generative Selector	**97.5**	81.9	**96.3**	**80.1**
Tu et al. (2020)	94.5	80.2	-	61.1
Fang et al. (2020)	-	**82.2**	-	78.9

Table 1: **HotpotQA:** Passage selection accuracy and end-to-end QA F1 on the original and adversarial set (Jiang and Bansal, 2019) of the HotpotQA dataset. The results of Tu et al. (2020) and Fang et al. (2020) are as reported by Perez et al. (2020).

Model	Accuracy	EM/F1
Standard Selector	96.8	72.8/79.9
Generative Selector	97.2	73.5/80.2

Table 2: **WikiHop:** Passage selection accuracy and end to end QA EM and F1 on dev set.

3.2 Context pairs vs. Sentences

Some context selection models for HotpotQA use a multi-label classifier that chooses top-k sentences (Fang et al., 2020; Clark and Gardner, 2018) which result in limited inter-document interaction than context pairs. To compare these two input types, we construct a multi-label sentence classifier $p(s|q, C)$ that selects relevant sentences. This classifier projects a concatenated sentence and question representation, followed by a sigmoid, to predict if the sentence should be selected. This model has a better performance over the context-pair selector but is more biased (Table 4).

We performed similar experiments with the generative model, where we train a generative *sentence* selection model by first selecting a set of sentences with a gumbel softmax (prior) and then generating

Context 1, c_i:	The America East Conference is a collegiate athletic conference affiliated with the NCAA Division I, whose members are located mainly in the Northeastern United States. The conference was known as the Eastern College Athletic Conference-North from 1979 to 1988 and the North Atlantic Conference from 1988 to 1996.
Context 2, c_j:	The Vermont Catamounts men's soccer team represents the University of Vermont in all NCAA Division I men's college soccer competitions. The team competes in the America East Conference.
Original Question, q:	the vermont catamounts men's soccer team currently competes in a conference that was formerly known as what from 1988 to 1996?
Generated Questions: $p(q \mid c_{ij}, \Psi)$	the vermont catamounts men's soccer team competes in what collegiate athletic conference affiliated with the ncaa division i, whose members are located mainly in the northeastern united states?
	the vermont catamounts men's soccer team competes in a conference that was known as what from 1979 to 1988?
	the vermont catamounts men's soccer team competes in a conference that was known as what from 1988 to 1996?

Table 3: Sample questions generated by using the question generation decoder with top-k sampling show that the generative model is able to construct (reason about) possible multi-hop questions given a context-pair, without lexically referencing the latent (bridge) entity,"American East Conference".

Model	Original	Adversarial
Discriminative Selectors		
Passage, $p(c_{ij}\|q, \Psi)$	95.3	96.3
Sentence, $p(s\|q, C)$	97.6	90.9
Generative Selectors		
Passage, $p(q \mid c_{ij}, \Psi)p(c_{ij}\|\Psi)$	97.5	96.3
Sentence, $p(q \mid s, C)p(s\|C)$	90.6	89.2
Multi-task, $p(q, s\|c_{ij}, \Psi)p(c_{ij}\|\Psi)$	98.1	97.2

Table 4: **Passages vs Sentences:** Passage selection accuracy for models with different context inputs on the development and adversarial set of HotpotQA.

the question given the set of sentences. Given that the space of set of sentences is much larger than context pairs, the generative sentence selector does not have good performance (Table 4).

Since sentence selection helped improve performance of the discriminative passage selector, we add an auxiliary loss term to our generative passage selector that also predicts the relevant sentences in the context pair when generating the question ($p(q, s|c_{ij}, \Psi)$), in a multi-task manner. We see slight performance improvements by using relevant sentences as an additional supervision signal.

4 Related work

Many recent passage selection models for HotpotQA and Wikihop's distractor style setup employ discriminative context selectors given the question (Tu et al., 2020; Fang et al., 2020; Shao et al., 2020). The high performance of such passage selectors can be attributed to existing bias in HotpotQA (Jiang and Bansal, 2019; Min et al., 2019b), which allows shallow lexical overlap of question with a single context to result in the correct answer. Another more general line of work dynamically updates the working memory to re-rank the set of passage at each hop (Das et al., 2019).

With the release of datasets like SearchQA (Dunn et al., 2017), TriviaQA (Joshi et al., 2017), and NaturalQuestions (Kwiatkowski et al., 2019), a lot of work has been done in open-domain passage retrieval, especially in the full Wikipedia setting. However, these questions do not necessarily require multi-hop reasoning. A series of work has tried to match a document-level summarized embedding to the question (Seo et al., 2018; Karpukhin et al., 2020; Lewis et al., 2020) for obtaining the relevant answers. In generative question answering, a few works (Lewis and Fan, 2019; Nogueira dos Santos et al., 2020) have used a joint question answering approach on a single context.

A large body of work has employed simple question generation for factoid answers in numerous cases, like answer verification (Duan et al., 2017), fact checking (Fan et al., 2020), data augmentation (Alberti et al., 2019; Serban et al., 2016), pedagogical systems (Lindberg et al., 2013), and dialog systems (Yanmeng et al., 2020) etc.

5 Conclusion

We proposed a generative formulation of context pair selection for multi-hop question answering. By encouraging this selection model to *explain* the entire question, it is less susceptible to bias, performing substantially better on adversarial data than existing discriminative methods. Our proposed model is simple to implement and can be used with *any* existing (or future) answering model; we

will release code to support this integration. Since context pair selection scales quadratically with the number of contexts, it is not ideal for scenarios that involve a large number of possible contexts. However, it allows for deeper inter-document interaction as compared to other approaches that use summarized document representations. With more reasoning steps, selecting relevant documents given only the question becomes challenging, increasing the need for inter-document interaction. An easy way to reduce the computation cost is to consider only a set of top-k contexts and perform a two-stage coarse-to-fine passage selection. The generative story presented in the paper may not work for question types beyond the datasets considered, for eg., in case of multiple (>1) bridge entities or more than two contexts. However, we demonstrate that this idea works for most common reasoning types that are central in current multi-hop reasoning datasets. The code is available at https://github.com/dDua/JointQA

6 Ethical Considerations

This paper focuses on how existing question answering models take shortcuts by exploiting biases in data that incentivize performing only shallow lexical overlap between the question and the context, to achieve better performance without learning the right way to reason. Based on the recommendations by Blodgett et al. (2020) (R3), if a systems is deployed in the wild, it should not favor inaccurate facts because of a brittle training methodology. Our work helps in taking a step towards understanding how to avoid exploiting such unwanted correlation in the data.

Acknowledgements

We would like to thank Robert Logan, Anthony Chen, Sanjay Subramanian and the anonymous reviewers for the discussions and feedback on earlier versions. We would also like to thank Hasso Plattner Institute (HPI) for supporting the first author through UCI-HPI fellowship.

References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, Florence, Italy. Association for Computational Linguistics.

Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. 2010. Impossibility theorems for domain adaptation. In *International Conference on Artificial Intelligence and Statistics*, pages 129–136.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Jifan Chen, Shih-ting Lin, and Greg Durrett. 2019. Multi-hop question answering via reasoning chains. *ArXiv preprint*, abs/1910.02610.

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, Melbourne, Australia. Association for Computational Linguistics.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. 2019. Multi-step retriever-reader interaction for scalable open-domain question answering. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Dheeru Dua, Sameer Singh, and Matt Gardner. 2020. Benefits of intermediate annotations in reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5627–5634, Online. Association for Computational Linguistics.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 866–874, Copenhagen, Denmark. Association for Computational Linguistics.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *ArXiv preprint*, abs/1704.05179.

Angela Fan, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel. 2020. Generating fact checking briefs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7147–7161, Online. Association for Computational Linguistics.

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2020. Hierarchical graph network for multi-hop question answering. In

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 8823–8838, Online. Association for Computational Linguistics.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. 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 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Yichen Jiang and Mohit Bansal. 2019. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop QA. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2726–2736, Florence, Italy. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Mike Lewis and Angela Fan. 2019. Generative question answering: Learning to answer the whole question. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

David Lindberg, Fred Popowich, John Nesbit, and Phil Winne. 2013. Generating natural language questions to support learning on-line. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 105–114, Sofia, Bulgaria. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Sewon Min, Danqi Chen, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019a. 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 2851–2864, Hong Kong, China. Association for Computational Linguistics.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019b. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4249–4257, Florence, Italy. Association for Computational Linguistics.

Cicero Nogueira dos Santos, Xiaofei Ma, Ramesh Nallapati, Zhiheng Huang, and Bing Xiang. 2020. Beyond [CLS] through ranking by generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1722–1727, Online. Association for Computational Linguistics.

Ethan Perez, Patrick Lewis, Wen-tau Yih, Kyunghyun Cho, and Douwe Kiela. 2020. Unsupervised question decomposition for question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*,

pages 8864–8880, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv preprint*, abs/1910.10683.

Minjoon Seo, Tom Kwiatkowski, Ankur Parikh, Ali Farhadi, and Hannaneh Hajishirzi. 2018. Phrase-indexed question answering: A new challenge for scalable document comprehension. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 559–564, Brussels, Belgium. Association for Computational Linguistics.

Iulian Vlad Serban, Alberto García-Durán, Caglar Gulcehre, Sungjin Ahn, Sarath Chandar, Aaron Courville, and Yoshua Bengio. 2016. Generating factoid questions with recurrent neural networks: The 30M factoid question-answer corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 588–598, Berlin, Germany. Association for Computational Linguistics.

Nan Shao, Yiming Cui, Ting Liu, Shijin Wang, and Guoping Hu. 2020. Is graph structure necessary for multi-hop reasoning? *ArXiv preprint*, abs/2004.03096.

Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. 2020. Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. In *AAAI*, pages 9073–9080.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Wang Yanmeng, Rong Wenge, Jianfei Zhang, Shijie Zhou, and Xiong Zhang. 2020. Multi-turn dialogue-oriented pretrained question generation model. *Complex & Intelligent Systems*, 6(3):493–505.