What is Missing in Existing Multi-hop Datasets? Toward Deeper Multi-hop Reasoning Task

Anonymous ACL submission

Abstract

Multi-hop machine reading comprehension (MRC) is a task that requires models to read and perform multi-hop reasoning over multiple paragraphs to answer a question. The task can be used to evaluate reasoning skills, as well as to check the explainability of the models, and is useful in applications (e.g., QA system). However, the current definition of *hop* (alias *step*) in the multi-hop MRC is ambiguous; moreover, previous studies demonstrated that many multi-hop examples contain reasoning shortcuts where the questions can be solved without performing multi-hop reasoning. In this opinion paper, we redefine multi-hop MRC to solve the ambiguity of its current definition by providing three different definitions of the steps. Inspired by the assessment of student learning in education, we introduce a new term of *In-depth multi-hop reasoning task* with three additional evaluations: step evaluation, coreference evaluation, and entity linking evaluation. In addition, we also examine the existing multi-hop datasets based on our proposed definitions. We observe that there is potential to extend the existing multi-hop datasets by including more intermediate evaluations to the task. To prevent reasoning shortcuts, multi-hop MRC datasets should focus more on providing a clear definition for the steps in the reasoning process and preparing gold data to evaluate them.

1 Introduction

The long-standing goal of natural language understanding (NLU) is to develop a machine that can understand natural languages like humans. Machine reading comprehension (MRC) is one of the most important tasks that can be used to evaluate NLU. MRC aims to teach computers to read and understand unstructured text automatically. In recent years, many datasets have been created, such as CNN/Daily Mail (Hermann et al., 2015) and SQuAD (Rajpurkar et al., 2016, 2018). Currently,

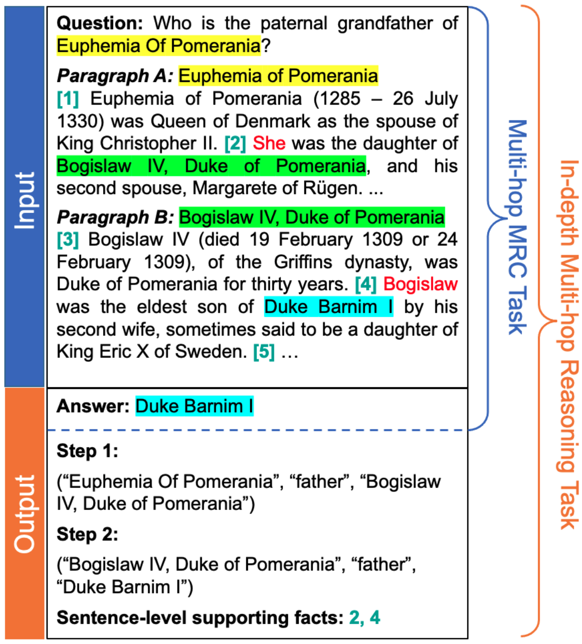

Figure 1: Examples of multi-hop MRC task and In-depth multi-hop reasoning task.

several models (Devlin et al., 2019; Liu et al., 2019; Yang et al., 2019) have outperformed humans (e.g., on SQuAD dataset). However, such performances do not indicate that these models can precisely understand the text. A major issue associated with these datasets is that they only provide a single paragraph as a context for a question, and questions are often answered via shallow lexical matching or based on various biases (Chen et al., 2016; Jia and Liang, 2017; Mudrakarta et al., 2018; Sugawara et al., 2018; Wang and Bansal, 2018).

Many attempts have been made to circumvent the issues described above, including unanswerable questions (Rajpurkar et al., 2018), knowledge-based MRC (Lai et al., 2017), conversational MRC (Reddy et al., 2019), and multi-hop MRC (Welbl et al., 2018). In this paper, we focus on the multi-hop MRC, which requires a model to answer a given question by reading and performing multi-hop reasoning over multiple paragraphs.

We argue that the current definition of multi-hop MRC is unclear. In particular, the definition of a hop (alias step) in the term multi-hop is ambiguous. Most of the previous datasets consider that the number of hops to be based on the number of paragraphs. This made the distinction between single-hop MRC and multi-hop MRC is vague. Owing to the rapid progress in the field, there are several multi-hop datasets that have been proposed for the task; however, previous studies demonstrated that many multi-hop samples do not require multi-hop reasoning to solve (Chen and Durrett, 2019; Jiang and Bansal, 2019; Min et al., 2019a; Trivedi et al., 2020). These samples contain reasoning shortcuts or some heuristic biases that models can use to answer the question.

Our goal in this opinion paper is to revise the current multi-hop MRC task and introduce the *In-depth multi-hop MRC task*. We first present the background and discuss the issues of the current definition of multi-hop MRC. To resolve those issues, we redefine the multi-hop MRC task. Inspired by the effects of intermediate assessment of student learning (Day et al., 2018), we introduce a new term *In-depth multi-hop reasoning task* (Figure 1) associated with three additional evaluations to comprehensively evaluate multi-hop models. We then examine the existing datasets based on our proposed definitions. Finally, we discuss some potential directions for future work on multi-hop MRC.

Given this redefinition and our proposal of *In-depth multi-hop reasoning task*, examining multi-hop datasets shows that most of the existing multi-hop datasets do not comprehensively explore the internal reasoning process from question to answer. We encourage future multi-hop datasets to focus extensively on the internal reasoning process and on preparing gold data to evaluate them.

2 Background

There are several existing tasks that require multi-hop reasoning, including multi-hop MRC (QA over text) (Welbl et al., 2018), QA over knowledge base (KB) (Zhang et al., 2018), QA over text and KB/tables (Chen et al., 2020), and claim verification (Jiang et al., 2020). In this paper, we focus on multi-hop MRC. We argue that the multi-hop MRC task is an important potential direction for the community in terms of the following attributes:

(i) Multi-hop MRC dataset is helpful for *testing the reasoning skills* of a model. To answer a multi-hop question, models must perform multiple reasoning steps. Each step often corresponds to several reasoning skills, such as comparisons and bridging entities.

(ii) Multi-hop MRC can be used to *evaluate the explainability* of a model. The internal reasoning process from a question to an answer involves multiple steps. Instead of evaluating models based solely on answer prediction task, previous studies (Yang et al., 2018; Ho et al., 2020; Inoue et al., 2020) have utilized internal reasoning information to evaluate the explainability of models.

(iii) Multi-hop MRC is useful in *applications*. Chen et al. (2017) introduced a way to construct a QA system by combining information retrieval (IR) and the MRC model. The MRC model in their system was designed for answering simple questions. However, questions in real-world QA systems can be complex and require many steps to be answered; Multi-hop MRC is an important component for answering those questions. Another application of multi-hop MRC is domain-specific information extraction, such as the discovery of drug-drug interactions by gathering information from different medical documents (Welbl et al., 2018).

To understand the issues related to the current definition of multi-hop MRC, we introduce the current definitions of single-hop MRC and multi-hop MRC in the next paragraph.

QA over Text (MRC): Single-hop MRC is defined as a task that requires a model to read one paragraph or document to answer a given question (Welbl et al., 2018; Yang et al., 2018). The task mainly focuses on testing the reasoning abilities of models in a single paragraph or document. In contrast, a multi-hop MRC task requires a model to read multiple paragraphs/documents to answer a question. Welbl et al. (2018) were the first to introduce the term *multi-hop reasoning*, and they also introduced the alias *multi-step reasoning*. They wanted to emphasize that instead of using only one document, the community should consider scenarios in which an answer is obtained by integrating information from multiple documents.

Current Issues: Based on the definitions above, we observe that there are two main issues associated with the current definition of multi-hop

MRC. (1) The first issue is the vagueness of the current definition; in particular, the distinction between the single-hop MRC and multi-hop MRC is unclear. When we concatenate multiple paragraphs/documents into one lengthy document, multi-hop questions become single-hop questions. (2) The second issue is about the reasoning shortcuts. Most previous multi-hop datasets have no evaluation to ensure that the models perform multi-hop reasoning. There can be shortcuts that make a question require fewer reasoning steps. Specifically, a previous work (Min et al., 2019a) demonstrated that multi-hop questions could become single-hop questions based on the information in distractor paragraphs (e.g., entity types).

3 Redefine Multi-hop MRC Task

To address the issues observed above, we redefine a multi-hop MRC task as follows:

Proposed Definition 1 *A multi-hop MRC task requires a model to perform "multiple steps" to answer questions.*

Owing to the diversity of multi-hop questions and the fact there are many ways to discover the internal reasoning processes from question to answer, we do not limit the definition of a step but only require a clear definition of steps and gold data for them. We introduce *three scenarios* with three different definitions of the steps in the path from question to answer. When using the following definitions of steps, the definition of the multi-hop MRC task is not based on the number of paragraphs.

Scenario 1 - A Step is a Sub-task: As discussed in previous works (Talmor and Berant, 2018; Min et al., 2019b), multi-hop questions can be decomposed into multiple simple sub-questions. For example, consider the question *Which team does the player named 2015 Diamond Head Classic's MVP play for?* We can split this question into two sub-questions: (a) *Which player was named 2015 Diamond Head Classic's MVP?* and (b) *Which team does ANS play for?* (*ANS* is the answer to the first sub-question). In this manner, we can consider predicting the answer to a sub-question as a step in the primary answering process.

Scenario 2 - A Step is a Triple: Previous works (Ho et al., 2020; Inoue et al., 2020) introduced a reasoning chain that describes relationships from the entities in the question to answer to explain the answers. Each triple in the reasoning chain can be considered as a step in the reasoning path from question to answer (Figure 1).

Scenario 3 - A Step is a Sequence of Tokens Containing a Single Operator: There are several works (Shi et al., 2020; Wolfson et al., 2020) on both MRC and QA over KB that have introduced an explicit reasoning process from question to answer. Wolfson et al. (2020) introduced a question decomposition meaning representation (QDMR) that contains a set of steps to find an answer. A step in QDMR is a sequence of tokens. Each step corresponds to a single query operator based on a set of predefined operators (e.g., group or sort). It is noteworthy that the QDMR is used as additional supervision data for training and not for evaluating the internal reasoning information (Figure 3).

4 In-depth Multi-hop Reasoning Task

Day et al. (2018) showed that from the teacher view, intermediate assessment could assess various knowledge and skills of students. Inspired by this finding, we introduce the new term *In-depth multi-hop reasoning task* consisting of three additional intermediate evaluations.

(i) **Step evaluation:** a dataset should provide a clear definition of the step and corresponding information that we can use to evaluate the model. This evaluation is essential because it can verify whether the model performs multiple steps when answering the question.

(ii) **Coreference resolution evaluation:** as discussed in Jurafsky and Martin (2020), coreference resolution (CR) is an important component of NLU. We observed that CR is important for multi-hop reasoning tasks. For example, in Figure 1, we cannot find the father of *Euphemia of Pomerania* if we do not know that the word "she" in the second sentence refers to *Euphemia of Pomerania*.

(iii) **Entity linking evaluation:** similar to coreference resolution evaluation, entity linking evaluation is necessary to verify the understanding of the model. For example, in Figure 1, we cannot find the father of *Bogislaw IV, Duke of Pomerania* if we do not know that the words "Bogislaw IV, Duke of Pomerania", "Bogislaw IV", and "Bogislaw" refer to the same person in paragraph B.

We argue that adding intermediate evaluations for multi-hop MRC task can prevent reasoning shortcuts. If the model performs reasoning shortcuts, then it cannot perform well on the intermediate tasks.

Dataset	Ans. style	Size	Corpus	Question source	Step scenario	Step evaluation	Core. & Ent. evaluation
QAngaroo (WikiHop) (Welbl et al., 2018)	MC	50K	Wikipedia	automated	✗	✗	✗
ComplexWebQues (Talmor and Berant, 2018)	Extr.	35K	Web snippet	automated & crowd	✗	✗	✗
HotpotQA (Yang et al., 2018)	Extr.	113K	Wikipedia	crowd	✓	✓	✗
OpenBookQA (Mihaylov et al., 2018)	MC	6K	textbook	crowd	✗	✗	✗
R^4C (Inoue et al., 2020)	Extr.	5K	Wikipedia	crowd	✓	✓	✗
2WikiMultiHopQA (Ho et al., 2020)	Extr.	200K	Wikipedia	automated	✓	✓	✗
HybridQA (Chen et al., 2020)	Extr.	70K	Wikipedia	crowd	✗	✗	✗
QASC (Khot et al., 2020)	MC	10K	textbook	crowd	✓	✗	✗
eQASC, eQASC-p, eOBQA (Jhamtani and Clark, 2020)	MC	10K	textbook	crowd	✓	✗	✗

Table 1: Existing multi-hop MRC datasets. For the column names: *Ans. style* represents answer style, *Step scenario* represents scenario 1/scenario 2/scenario 3 (Section 3), *Core. & Ent. evaluation* represent coreference resolution evaluation and entity linking evaluation. In the *Ans. style* column, "Extr." represents extraction and "MC" denotes multiple-choice.

5 Examine Multi-hop Datasets

We present all the existing multi-hop datasets (Appendix C) in Table 1. It should be noted that our focus is not to compare existing multi-hop datasets; instead, we want to provide an overview for the community via this table. We can observe that most datasets have not explored the details of the internal reasoning process from question to answer; specifically, there are only two datasets (datasets with a green check) that have been provided to evaluate the internal reasoning process. Recently, Tang et al. (2021) introduced an additional sub-question evaluation (the blue check) for HotpotQA. However, the authors provided only 1,000 sub-questions for the evaluation. Instead of focusing on outside of the reasoning process, such as constructing adversarial paragraphs (Jiang and Bansal, 2019) or using a single-hop model (Min et al., 2019a), we suggest that the community should focus on the internal reasoning process by providing and successively evaluating all information in the reasoning path.

6 Discussion & Conclusion

In this section, we first discuss some directions for future work on multi-hop MRC and then conclude our paper. We observe that there are various directions for improving multi-hop datasets. The first is about explainability. Instead of focusing on model explainability, we can shift the focus to dataset explainability (Sugawara et al., 2021). Multi-hop questions contain many steps in their internal reasoning processes, from a question to an answer. Therefore, evaluating the models successively on the path from question to answer is an effective way of testing the explainability of models.

The second is about reasoning skills. Multi-hop questions can potentially require diverse reasoning skills (e.g., comparisons and bridging entities) to arrive at an answer. However, currently, there are no multi-hop datasets that provide the reasoning skills required for answering questions. There has therefore been no analysis on which reasoning skills are more difficult for models and which reasoning skills models perform well on. We argue that incorporating a set of skills (Sugawara et al., 2017) for each sample in a multi-hop dataset is an effective method for evaluating and improving multi-hop models.

In conclusion, in this paper, we redefined the multi-hop MRC task and provided a new definition of a *In-depth multi-hop reasoning task* for comprehensively evaluating multi-hop models. We also examined the existing datasets based on our proposed definitions, and finally, we discussed several directions for future work on multi-hop MRC tasks.

References

Danqi Chen, Jason Bolton, and Christopher D. Manning. 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. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. 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. Association for Computational Linguistics.

Jifan Chen and Greg Durrett. 2019. Understanding dataset design choices for multi-hop reasoning. 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 4026–4032, Minneapolis, Minnesota. Association for Computational Linguistics.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online. Association for Computational Linguistics.

Indira N. Z. Day, F. M. van Blankenstein, P. Michiel Westenberg, and W. F. Admiraal. 2018. Teacher and student perceptions of intermediate assessment in higher education. *Educational Studies*, 44(4):449–467.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1693–1701, Cambridge, MA, USA. MIT Press.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Naoya Inoue, Pontus Stenetorp, and Kentaro Inui. 2020. R4C: A benchmark for evaluating RC systems to get the right answer for the right reason. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6740–6750, Online. Association for Computational Linguistics.

Harsh Jhamtani and Peter Clark. 2020. Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 137–150, Online. 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.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2020. *Speech and Language Processing*. USA.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. *AAAI*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 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. 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*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019a. 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.

Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019b. Multi-hop reading comprehension through question decomposition and rescoring. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6097–6109, Florence, Italy. Association for Computational Linguistics.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the model understand the question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1896–1906, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 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. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 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. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Jiaxin Shi, Shulin Cao, Liangming Pan, Yutong Xiang, Lei Hou, Juanzi Li, Hanwang Zhang, and Bin He. 2020. Kqa pro: A large-scale dataset with interpretable programs and accurate sparqls for complex question answering over knowledge base. *arXiv*.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4219, Brussels, Belgium. Association for Computational Linguistics.

Saku Sugawara, Yusuke Kido, Hikaru Yokono, and Akiko Aizawa. 2017. Evaluation metrics for machine reading comprehension: Prerequisite skills and readability. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 806–817, Vancouver, Canada. Association for Computational Linguistics.

Saku Sugawara, Pontus Stenetorp, and Akiko Aizawa. 2021. Benchmarking machine reading comprehension: A psychological perspective. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1592–1612, Online. Association for Computational Linguistics.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

Yixuan Tang, Hwee Tou Ng, and Anthony Tung. 2021. Do multi-hop question answering systems know how to answer the single-hop sub-questions? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3244–3249, Online. Association for Computational Linguistics.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2020. Is multihop QA in DiRe condition? measuring and reducing disconnected reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8846–8863, Online. Association for Computational Linguistics.

Yicheng Wang and Mohit Bansal. 2018. Robust machine comprehension models via adversarial training. 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 575–581, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Tomer Wolfson, Mor Geva, Ankit Gupta, Matt Gardner, Yoav Goldberg, Daniel Deutch, and Jonathan Berant. 2020. Break it down: A question understanding benchmark. *Transactions of the Association for Computational Linguistics*, 8:183–198.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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.

Wen-tau Yih, Matthew Richardson, Chris Meek, Ming-Wei Chang, and Jina Suh. 2016. The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–206, Berlin, Germany. Association for Computational Linguistics.

Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander J. Smola, and Le Song. 2018. Variational reasoning for question answering with knowledge graph. In *AAAI*.

A Redefine Multi-hop MRC Task — Details

We argue that most complex/compositional questions are multi-hop questions; however, based on the provided documents, a model can use various heuristics or simple rules that we are unaware of to answer a question. For example, there may be a question that asks about an animal in ten paragraphs (considered as supporting documents) but with only one paragraph about the animal. In this case, the question can be a single-hop question; therefore, we should design datasets with multiple tasks by using intermediate information instead of only having an answer prediction task.

Scenario 2 - A Step is a Triple: Figure 2 illustrates a multi-hop question where a step is a triple. The question in this example is called comparison question. This type of question is introduced in HotpotQA and 2WikiMultiHopQA. We argue that when the question type is a comparison question, a set of triples is not enough to explain the answer. In this example, we can obtain the two triples about the date of birth of *George Washington* or *Martha Washington*. However, to obtain the final answer, we need to perform one more step is to compare the two dates: *February 22, 1732* and *June 13, 1731*.

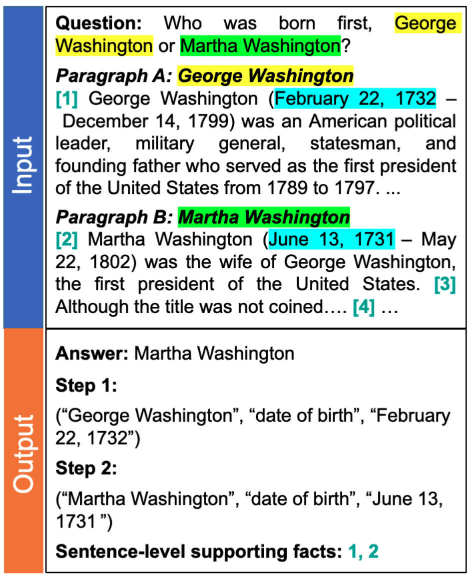

Figure 2: An example of a multi-hop question where a step is a triple (scenario 2).

Scenario 3 - A Step is a Sequence of Tokens Containing a Single Operator: Figure 3 illustrates an example of a multi-hop question with QDMR.

We observe that QDMR is a promising information to represent the reasoning process from question to answer. However, the current research does not utilize this information for evaluation.

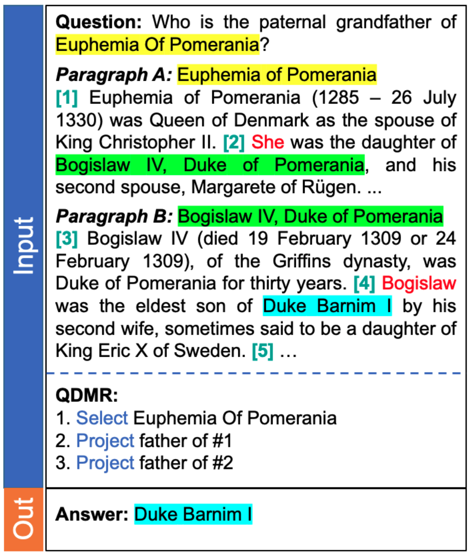

Figure 3: An example of a multi-hop question with QDMR.

B In-depth Multi-hop Reasoning Task — Details

In this appendix, we propose a way to implement coreference resolution and entity linking evaluations for multi-hop MRC datasets. We do not apply these evaluations for all entities in the context. Instead, we focus on entities related to the reasoning path from the entity in the question that leads to the answer.

Coreference Resolution Evaluation: For each entity in the reasoning path, this evaluation requires a model to predict all pronouns that refer to the entity from where the entity starts until the end of the triple corresponding to the entity. For example, in Figure 1, the ground truth labels for all entities are:

- Euphemia of Pomerania: {She}
- Bogislaw IV, Duke of Pomerania: { }

Entity Linking Evaluation: In contrast to coreference resolution evaluation, this evaluation requires a model to predict all other entity names that refer to the entity from where the entity starts until the answer. For example, in Figure 1, the ground truth labels for all entities are:

- Euphemia of Pomerania: {}
- Bogislaw IV, Duke of Pomerania: {"Bogislaw IV", "Bogislaw"}

C Existing Multi-hop Datasets

QAngaroo (Welbl et al., 2018) was the first dataset where multi-hop reasoning in MRC was introduced. This dataset contains two sub-datasets called WikiHop and MedHop in the open domain and medicine domain, respectively. The dataset was constructed based on KB and Wikipedia. Subsequently, Talmor and Berant (2018) introduced ComplexWebQuestions, a dataset created by making the WebQuestionSP dataset (Yih et al., 2016) more complicated. Owing to their building procedures, both datasets do not provide any information to explain the predicted answers. Later, Yang et al. (2018) introduced HotpotQA, a crowdsourced dataset. In HotpotQA, the authors introduced new information called *sentence-level supporting facts*, which are sets of sentences that support answers. They also introduced a new task called sentence-level supporting fact prediction, which is a binary classification task. This type of explanation is called a justification explanation (collection of evidence to support a decision). Subsequently, Inoue et al. (2020) introduced a new dataset called $\mathrm{R}^4\mathrm{C}$ that provides both justification and introspective explanations (how a decision is made). Following that direction, Ho et al. (2020) introduced the 2WikiMultiHopQA dataset, which was constructed by utilizing KB and Wikipedia. The difference between $\mathrm{R}^4\mathrm{C}$ and 2WikiMultiHopQA lies in the manner in which they represent introspective explanation information, where the former uses semi-structured data and the latter uses structured data. Additionally, the targets of $\mathrm{R}^4\mathrm{C}$ and 2WikiMultiHopQA are also different: $\mathrm{R}^4\mathrm{C}$ focuses on the internal reasoning process (it was created based on HotpotQA and only contains 4,588 questions); in contrast, 2WikiMultiHopQA was designed to focus on the entire reasoning process from question to answer.

In addition to the datasets discussed above, there is another dataset that requires multi-hop reasoning for both structured and unstructured text. Recently, Chen et al. (2020) introduced the HybridQA dataset, which requires reasoning over both tabular and textual data to answer questions. This dataset was created by crowdsourcing based on Wikipedia tables and articles. There are three main steps: table/passage collection, question/answer collection, and annotation de-biasing. To ease for annotators and ensure the quality of the dataset, the authors use some rules in the dataset collection process, such as choosing tables with rows between 5-20 and restraining tables from having many hyperlinked cells.

In contrast to the datasets discussed above, Mihaylov et al. (2018) introduced the OpenBookQA dataset, which requires multi-hop reasoning and combines open book facts with additional common knowledge facts (from external sources) to answer multiple-choice questions. A notable feature of this dataset is that the questions do not contain sufficient information to decompose them into multiple facts/sub-questions. However, it is unclear how many additional facts are required, whether models must use additional facts, or whether facts are available from external common knowledge sources. To address these issues, Khot et al. (2020) introduced QASC, which is a multi-hop reasoning dataset based on sentence composition that focuses on fact compositions. They explicitly identified two facts that were required to answer a target question. The two facts were created by crowdsourcing. However, QASC only provides one explanation for each question-answer pair. In reality, there may be a number of valid explanations. To tackle this issue, Jhamtani and Clark (2020) introduced three explanation datasets called eQASC, eQASC-perturbed, and eOBQA, which were created by reusing QASC and OpenBookQA.