How Well Do Multi-hop Reading Comprehension Models Understand Date Information?

Anonymous ACL submission

Abstract

Many previous works demonstrated that existing multi-hop reading comprehension datasets (e.g., HotpotQA) contain reasoning shortcuts, where the questions can be answered without performing multi-hop reasoning. Recently, several multi-hop datasets have been proposed to solve the reasoning shortcut problem or evaluate the internal reasoning process. However, the design of the reasoning chain for comparison questions in R^4C and 2WikiMultiHopQA does not fully explain the answer; meanwhile, MuSiQue only focuses on bridge questions. Therefore, it is unclear about the ability of a model to perform step-by-step reasoning when finding an answer for a comparison question that requires comparison and numerical reasoning skills. To evaluate the model completely in a hierarchical manner, we first propose a dataset, *HieraDate*, created by reusing and enhancing two previous multi-hop datasets, HotpotQA and 2WikiMultiHopQA. Our dataset focuses on comparison questions on date information that require multi-hop reasoning for solving. We then evaluate the ability of existing models to understand date at three levels: extraction, reasoning, and robustness. Our experimental results reveal that the multi-hop models fail at the reasoning level. Comparison reasoning and numerical reasoning (e.g., subtraction) are key challenges that need to be addressed in future works.

1 Introduction

Multi-hop machine reading comprehension (MRC) requires a model to read and aggregate information from multiple paragraphs to answer a given question (Welbl et al., 2018). Several datasets have been proposed for the task, such as HotpotQA (Yang et al., 2018) and 2WikiMultiHopQA[1] (Ho et al., 2020). Although the proposed models show promising performances, previous studies (Jiang and Bansal, 2019; Chen and Durrett, 2019; Min et al., 2019a; Tang et al., 2021) have demonstrated that existing multi-hop datasets contain reasoning shortcuts, in which case the model can answer the question without performing multi-hop reasoning.

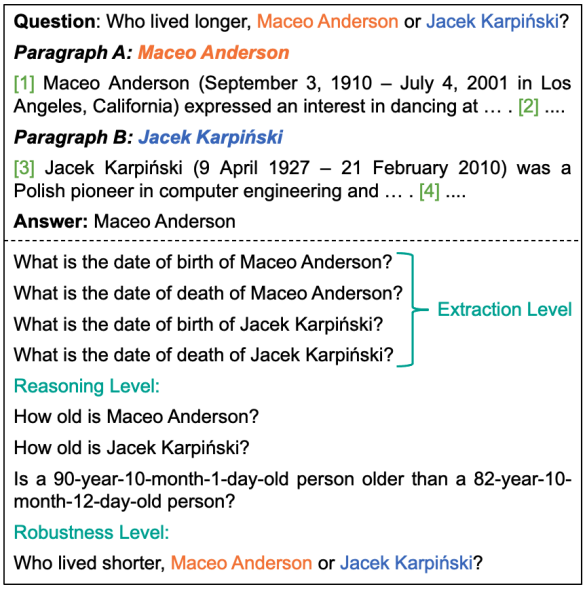

Figure 1: Example of a question in our dataset.

Specifically, using adversarial examples, Jiang and Bansal (2019) showed that multi-hop questions can be solved by matching the words in a question with a sentence in context. Chen and Durrett (2019) and Min et al. (2019a) designed a sentence-factored model and a single-hop BERT-based model, respectively, to test the necessity for multi-hop reasoning. By design, these models did not possess the ability to answer multi-hop questions; however, the results indicated that they could answer a large number of examples. Although these studies demonstrate the existence of reasoning shortcuts, they do not clarify in detail the internal reasoning processes of the question-answering (QA) process.

In general, there are two main types of questions in previous multi-hop datasets: bridge and comparison. Tang et al. (2021) explored the sub-questions in the QA process for model evaluation. However,

[1]For brevity, we use 2Wiki to denote 2WikiMultiHopQA.

they only used the bridge questions in HotpotQA. Min et al. (2019a) showed that comparison questions have fewer reasoning shortcuts than bridge questions. We argue that the multi-hop reasoning ability of a model remains unclear when evaluated on examples that do not ensure the requirement of multi-hop reasoning. On the other hand, based on the classification of comparison questions in Min et al. (2019a) and through manual analysis, we observe that most comparison questions on date information require multi-hop reasoning for solving.

HotpotQA only provides sentence-level supporting facts (SFs) to explain the answer, whereas 2Wiki provides both sentence-level SFs and evidence information. The evidence is a set of triples. For example, for the question in Figure 1, the evidence is about the date of birth and date of death of two people, e.g., *(Maceo Anderson, date of birth, September 3, 1910)*. We argue that simply requiring the models to detect a set of triples, in this case, cannot explain the answer for the question and cannot describe the full path from question to answer; additional operations including calculations and comparisons need to be performed to obtain the final answer.

Motivated by the example in Figure 1, we organize the sub-questions and adversarial questions into three levels: *extraction*, *reasoning*, and *robustness*, to evaluate of the ability of the existing multi-hop models to understand date information. Figure 1 depicts an example of a comparison question in our dataset. In this study, we aim to answer the following questions: (1) *Extraction level*: Does the model know how to obtain the date information (e.g., date of birth)? (2) *Reasoning level*: Does the model know how to calculate the age given the two dates (date of birth and death)? Can the model compare two dates or two ages to obtain the answer? (3) *Robustness level*: Can the model answer correctly if the question is slightly modified (e.g., flip the answer)?

We first propose a dataset, *HieraDate*, created by reusing and enhancing two existing multi-hop datasets, HotpotQA and 2Wiki. As the first step of the proof of concept, we focus on understanding the date information through comparison questions because this information is easy to handle and control. Moreover, our comparison questions on the date information ensure that multi-hop reasoning is required. We then evaluate the ability of existing multi-hop models to understand the date information in our dataset. Additionally, we obtain human performance on 100 random samples in our dataset.

Experimental results reveal that the existing multi-hop models can perform well at the extraction and robustness levels but completely fail at the reasoning level, including both comparison and numerical reasoning. The predicted answers are incorrect for 88.3% and 89.7% of the samples, for comparison questions and numerical questions, respectively, although the corresponding main multi-hop questions are answered correctly. These results demonstrate that the main issue in existing multi-hop models is the deficiency in comparison reasoning and numerical reasoning. Moreover, the human annotation scores indicate that our dataset is sufficiently good for evaluation.

2 Related Work

There are several multi-hop datasets, including those constructed using knowledge bases (e.g., WikiHop (Welbl et al., 2018) and ComplexWebQuestion (Talmor and Berant, 2018)) and those created through crowdsourcing (e.g., HotpotQA (Yang et al., 2018) and $\mathrm{R}^4\mathrm{C}$ (Inoue et al., 2020)). MuSiQue (Trivedi et al., 2021) was proposed recently, although their experiments showed that MuSiQue is less cheatable than HotpotQA and 2Wiki, it contains only bridge questions. Wolfson et al. (2020) proposed a Break dataset with the question decomposition meaning representation (QDMR) information. They annotate QDMR for samples in HotpotQA. Our dataset is different from Break, where we have both sub-questions and sub-answers; meanwhile, they only provide a list of steps (sub-questions).

There are several datasets related to numerical reasoning, such as DROP (Dua et al., 2019) and NOAHQA (Zhang et al., 2021). Our dataset is different from these datasets, to the best of our knowledge, our dataset is the first dataset that combines both numerical reasoning and multi-hop reasoning.

There are two previous works (Tang et al., 2021; Al-Negheimish et al., 2021) that are similar to ours. Tang et al. (2021) evaluated the previous models on the bridge questions in HotpotQA. Our work differs because we focus on comparison questions and organize the information in a hierarchical manner. Al-Negheimish et al. (2021) evaluated the previous models on the DROP dataset to test their numerical reasoning ability. However, they did not investigate the internal reasoning processes of these models.

Split	Main	Extraction	Comparison Reasoning	Combined Reasoning	Robustness	Total
dev	549	1346	425 (x2)	124 (x3)	549	3666
test	549	1346	425 (x2)	124 (x3)	549	3666
train	8745	21340	6820 (x2)	1925 (x3)	8745	58245

Table 1: Our dataset information.

3 Dataset Construction

We briefly describe the two existing multi-hop datasets, HotpotQA and 2Wiki. We then describe the generation of our dataset.

HotpotQA (Yang et al., 2018): HotpotQA, created through crowdsourcing, includes two main types of questions: bridge and comparison. Unlike previous datasets, a set of sentence-level SFs information is introduced in HotpotQA, which facilitates explainable reasoning by the system. Because of the dataset construction procedure, there is no available information in HotpotQA that can be used to generate sub-questions.

2WikiMultiHopQA (Ho et al., 2020): 2Wiki was created using Wikipedia articles and Wikidata triples. Similar to HotpotQA, it includes two main types of questions: bridge and comparison. In 2Wiki, the authors introduced evidence information that can be used to explain the reasoning chain from question to answer. We used this information for generating sub-questions in our dataset.

Our Dataset: We first filtered the samples by selecting only the comparison questions in HotpotQA and 2Wiki. We then used a set of keywords, such as *born first* and *lived longer*, to obtain a set of questions on the date information. For dev and train, respectively, we obtained 114 and 878 samples in HotpotQA, and 984 and 8745 samples in 2Wiki.

In 2Wiki, we used the evidence and Wikidata IDs (available in 2Wiki) to automatically generate sub-questions and sub-answers for two levels: extraction and reasoning. We observed that nine phrases (e.g., *born first*) could cover all the questions, and used these phrases to generate the robustness questions. In HotpotQA, we first filtered the distractor paragraphs and retained only two gold paragraphs for annotation. We then used Spacy to extract the entities in the questions. Further, we manually annotated the date with two formats: *string* and *dictionary*. Finally, we processed the annotated samples to generate all questions at three levels: extraction, reasoning, and robustness. It is noted that we used only the dev set in HotpotQA.

We combined the dev sets in HotpotQA and 2Wiki to create the dev and test sets in our dataset. We used the train set in 2Wiki as the training set of our dataset. Table 1 lists the number of samples at each level and each subset in our dataset. Our dataset includes two main types of questions: questions that require both date-of-birth and date-of-death information (e.g., "who lived longer"), and those that require only the date-of-birth or date-of-death information (e.g., "who was born first"). We call the first type *combined reasoning* because it requires both types of reasoning: comparison and numerical (Figure 1). The second type is called *comparison reasoning* (Figure 2) because it requires only comparison reasoning.

4 Experiments

We first briefly introduce the top-performing models used in our experiments. Further, we present the results of these models and the human performance results. Finally, we analyze the results.

Models: As the existing models cannot perform on all the three levels, we evaluate these models as two groups: one focusing on comparison reasoning (e.g., the SAE and HGN models) and the other focusing on numerical reasoning (e.g., the NumNet+ model). SAE (Tu et al., 2020) and HGN (Fang et al., 2020) were designed to deal with HotpotQA, whereas NumNet+ (Ran et al., 2019) was designed to deal with DROP (Dua et al., 2019). All these three models can perform on the extraction as well as robustness levels. By design, SAE and HGN can address yes/no questions; therefore, they can perform on yes/no questions in the comparison reasoning level. Meanwhile, NumNet+ cannot address yes/no questions, but can deal with numerical reasoning questions (Appendix B.1).

Results: Table 2 presents the results of the existing models on the test set of our dataset. The results indicate that both SAE and HGN perform well at

Model	Main		Extraction		Reasoning		Robustness	
	EM	F1	EM	F1	EM (num)	EM (compare)	EM	F1
SAE	69.76	77.78	82.99	84.73	×	59.14	69.22	77.82
HGN	66.85	76.15	94.58	96.14	×	52.98	71.95	81.64
NumNet+	67.94	71.57	1.26	47.93	8.21 (F1)	×	69.58	71.91
Human (average)	94.00	94.90	99.16	99.53	100	98.06	95.5	95.9
Human UB	100	100	100	100	100	100	100	100

Table 2: Results (%) of the previous models on the test set of our dataset. *Num* denotes numerical reasoning and *compare* denotes comparison reasoning. *Human UB* represents the human upper bound.

the extraction level, but their results for comparison reasoning are not as good. As all questions in the comparison reasoning level are yes/no questions, the random score is 50%. In this case, the HGN score is close to the chance score. At the robustness level, the results are comparable with those of the main multi-hop questions. This can be explained by the fact that the patterns of the main multi-hop questions and robustness questions are similar.

The results of the main multi-hop questions and robustness questions are comparable for the NumNet+ model as well. However, the model fails completely at the extraction and reasoning levels. At the extraction level, the model could not predict the complete span of the date, including the year, month, and day, but could predict the year. For numerical reasoning, there are two methods to convert our dataset to the format of the DROP dataset when dealing with the age. (1) The age is retained at three levels: year, month, and day; we refer to this as the *date* format. (2) Only the year of the age is retained; we refer to this as the *number* format. Through experiments, we observed that NumNet+ could not subtract two dates, but could subtract two numbers. In Table 2, we present the results of the date format. When the dataset is converted to the number format, the EM score on the numerical reasoning level is 21.8%.

Human Performance: We randomly selected 100 samples from the test set for human annotation. Each sample was annotated by two graduate students. We provided the context and a list of questions to the students; the results are depicted in Table 2. It can be observed that the human upper bound is 100% for all the scores. However, the human average is slightly low. On manually investigating the reason for this low human average, we found that the students made mistakes in several examples, which are answerable and reasonable. This indicates that our dataset is sufficiently good for evaluation.

Analyses: We analyzed each level separately to determine the capability of the models. We examined the number of cases where the predicted main multi-hop answer was correct, but the predicted answer for each level was incorrect. Table 4 depicts the number of samples where the sub-questions are incorrectly answered, but the predicted main multi-hop answer is correct. In 89.7% of the samples, the numerical question was answered incorrectly, but the main multi-hop question was correctly answered; for comparison questions, it is 88.3%. These results indicate that the models completely fail in comparison and numerical reasoning when answering date questions. Although there are only 28.2% incorrectly answered robustness questions when the predicted main multi-hop answer is correct, it does not prove the multi-hop reasoning capability of the model because the robustness level depends on the extraction and reasoning levels.

5 Conclusion

We proposed a new dataset for comprehensively evaluating the ability of existing models to understand date information. We evaluated the top-performing multi-hop models on our dataset. Experimental results and analyses revealed that these models could not perform numerical reasoning and comparison reasoning, although the corresponding multi-hop questions were correctly answered.

For future work, we intend to use model explanation techniques and build a model that performs all the reasoning steps to discover which features that the model uses to answer the questions. We also intend to use the hierarchical manner in our dataset to apply for other types of questions.

References

Hadeel Al-Negheimish, Pranava Madhyastha, and Alessandra Russo. 2021. Numerical reasoning in machine reading comprehension tasks: are we there yet? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9643–9649, Online and Punta Cana, Dominican Republic. 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.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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 2368–2378, Minneapolis, Minnesota. 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.

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.

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.

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.

Qiu Ran, Yankai Lin, Peng Li, Jie Zhou, and Zhiyuan Liu. 2019. NumNet: Machine reading comprehension with numerical reasoning. 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 2474–2484, Hong Kong, China. 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. 2021. Musique: Multi-hop questions via single-hop question composition. *arXiv:2108.00573*.

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 2020*.

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, 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.

Qiyuan Zhang, Lei Wang, Sicheng Yu, Shuohang Wang, Yang Wang, Jing Jiang, and Ee-Peng Lim. 2021. NOAHQA: Numerical reasoning with interpretable graph question answering dataset. *arXiv*.

A Dataset Details

Date Format: Wikidata uses a zero value for the dates that miss the month value or day value. In reality, we have no date with month-0 and day-0; therefore, we use a default value (1) for the dates that miss the month value or day value.

Question Types: As mentioned above, there are two main types of questions in our dataset: combined reasoning (Figure 1) and comparison reasoning (Figure 2). One combined reasoning sample has one main multi-hop question, four extraction questions, two numerical reasoning questions, one comparison question, and one robustness question. Meanwhile, one comparison reasoning sample has one main multi-hop question, two extraction questions, two comparison questions, and one robustness question.

After obtaining all samples from HotpotQA and 2Wiki, there are only 11.3% of combined questions in the total number of examples. Therefore, we use some rules and heuristics to automatically convert samples that satisfy the requirements to become combined reasoning questions. In the current version of the dataset, there are 22.1% of combined questions.

Figure 2: Example of a question in our dataset.

Numerical Reasoning Issue: In reality, in some cases, the paragraph can contain age information, e.g., "*He died in 1981 at the age of 90*". In this case, the model does not need to perform numerical reasoning. We used rules, then manually checked, and found that there are 13 paragraphs in a total of 248 paragraphs (124 examples) in the test set that the age information is available.

Dataset Versions: We released two versions of the dataset: one with a "normal setting" including only two gold paragraphs that serve as the context, and the other with a "distractor setting" having ten paragraphs including two gold paragraphs and eight distractor paragraphs. In this study, we evaluated the previous models on the "normal setting" version.

B Experiments

B.1 Model

NumNet: There are some versions of the NumNet model; in our experiment, we use the NumNet+ version[2]. There are two ways to convert the extraction level questions in our dataset to the format of the DROP dataset. One is to use a span format, and another one is to use a date format; in our experiment, we use the span format.

B.2 Evaluation on DecompRC

We also evaluate a question decomposition system (DecompRC (Min et al., 2019b)) in our dataset. We use the original code[3] to evaluate the comparison questions in our dataset. In DecompRC, the multi-hop question is decomposed into two sub-questions and one operator question by using heuristics and rules. However, the heuristics and rules cannot cover all cases where the name entities are ambiguous. We use DecompRC to decompose 549 samples on the test set of our dataset, but it can decompose only 504 samples. We evaluate DecompRC on these 504 samples.

We do not evaluate the DecompRC system at the reasoning level because DecompRC only performs rules to obtain the final answer for comparison reasoning and numerical reasoning questions. The results are presented in Table 3. The main reason why the scores of the main multi-hop questions and the robustness questions are low is the predicted operator. The predicted operator is very important for the final answer, but many predicted operators are incorrect. For example, for the question "*Who was born later, person A or person B?*", the operator should be "*Which is greater*", but the DecompRC system often predicts "*Which is true*".

[2] https://github.com/llamazing/numnet_plus

[3] https://github.com/shmsw25/DecompRC

Model	Main		Extraction		Reasoning		Robustness	
	EM	F1	EM	F1	EM (num)	EM (compare)	EM	F1
DecompRC	44.64	47.26	82.60	83.94	×	×	38.89	40.16

Table 3: Results (%) of the DecompRC system on the test set of our dataset. *Num* denotes numerical reasoning and *compare* denotes comparison reasoning.

Model	Level	#Samples	Incorrect_sub/correct_main	%
SAE	Extraction	549	99 / 383	25.85
	Reasoning (compare)	549	262 / 383	68.41
	Robustness	549	108 / 383	**28.20**
HGN	Extraction	549	17 / 367	4.63
	Reasoning (compare)	549	324 / 367	**88.28**
	Robustness	549	96 / 367	26.16
NumNet+	Extraction	549	372 / 373	99.73
	Reasoning (num)	124	61 / 68	**89.71**
	Robustness	549	66 / 373	17.69

Table 4: Number of samples where the sub-questions are incorrectly answered, but the predicted main multi-hop answer is correct.

B.3 Analyses

Table 4 presents the information of our analysis for three levels of the three models. One sample can have two sub-questions in the comparison reasoning level or the numerical reasoning level. Suppose one of two sub-questions is incorrect and the predicted main multi-hop answer is correct; in that case, we count the sample into the total number of samples where the sub-questions are incorrectly answered, but the predicted main multi-hop answer is correct.

On the leaderboard of HotpotQA[4], SAE performs better than HGN. We also observe this behavior in our dataset. In HGN, there are 88.3% of samples where the comparison question is incorrect, but the main multi-hop question is correct. However, the percentage drops in SAE, it is only 68.4%. One interesting point is that there are only 4.6% of samples where the extraction question is incorrect, but the main multi-hop question is correct in HGN. However, the percentage is 25.8% in SAE. This indicates that HGN is better than SAE at the extraction level.

Table 5 presents some error cases of the previous models on the test set of our dataset. In the first two examples, we can see that the models do not have the ability to compare two dates. In examples #3 and #4, we can observe that the models do not have the ability to calculate the age from the two dates. In some cases, the models can calculate the age by simplifying subtract two years of the two dates. In example #5, we observe that the models can answer the main multi-hop question correctly, although they do not know what the date of death of a person is.

[4] https://hotpotqa.github.io/

Context	Main question	Sub-questions
Paragraph A: Lotte Backes (May 2, 1901 - May 12, 1990) was a German pianist, ... **Paragraph B:** Willem van Haecht (1593 – 12 July 1637) was a Flemish painter best known for his pictures ...	**Q:** Who died first, Lotte Backes or Willem van Haecht? **Predicted answer:** Willem van Haecht ✓	**Q1:** Does May 12, 1990 come before July 12, 1637? **Predicted 1:** yes ✗ **Q2:** Does May 12, 1990 come after July 12, 1637? **Predicted 2:** yes ✓
Paragraph A: Taiwo Afolabi (born April 29, 1962), is a Nigerian business magnate ... **Paragraph B:** Claudio O'Connor (born April 28, 1963 in Llavallol, Buenos Aires Province) is a thrash metal ...	**Q:** Who was born later, Taiwo Afolabi or Claudio O'Connor? **Predicted answer:** Claudio O'Connor ✓	**Q1:** Does April 29, 1962 come before April 28, 1963? **Predicted 1:** no ✗ **Q2:** Does April 29, 1962 come after April 28, 1963? **Predicted 2:** no ✓
Paragraph A: Andrzej Markowski (22 August 1924 – 30 October 1986) was a Polish composer and conductor. ... **Paragraph B:** François Missoffe (13 October 1919 in Toulon, France – 28 August 2003 in Rouen) was a French politician and diplomat. ...	**Q:** Who lived longer, Andrzej Markowski or François Missoffe? **Predicted answer:** François Missoffe ✓	**Q1:** How old is Andrzej Markowski? **Predicted 1:** 1924 ✗ **Q2:** How old is François Missoffe? **Predicted 2:** 1919 ✗ **Q3:** Is a 62-year-2-month-8-day-old person older than a 83-year-10-month-15-day-old person? **Predicted 3:** yes ✗
Paragraph A: Dyson Carter (February 2, 1910 – 1996) was a Canadian scientist, lecturer, writer, ... **Paragraph B:** Arne Kotte (20 March 1935 – 8 July 2015) was a Norwegian footballer who played as a forward. ...	**Q:** Who lived longer, Dyson Carter or Arne Kotte? **Predicted answer:** Dyson Carter ✓	**Q1:** How old is Dyson Carter? **Predicted 1:** 1910 ✗ **Q2:** How old is Arne Kotte? **Predicted 2:** 80 (number format) ✓ **Q3:** Is a 85-year-10-month-30-day-old person younger than a 80-year-3-month-18-day-old person? **Predicted 3:** yes ✗
Paragraph A: Oliver A. Unger (August 28, 1914 – March 27, 1981) was an award- winning American film producer, distributor, ... **Paragraph B:** Ross Story (16 January 1920 – 9 May 1991), always known as Ross or C. R. Story, was a farmer and politician ...	**Q:** Who died later, Oliver A. Unger or Ross Story? **Predicted answer:** Ross Story ✓	**Q1:** What is the death date of Oliver A. Unger? **Predicted 1:** 9 May 1991 ✗ **Q2:** What's the death date of Ross Story? **Predicted 2:** 9 May 1991 ✓ **Q3:** Does March 27, 1981 come before May 09, 1991? **Predicted 3:** yes ✓ **Q4:** Does March 27, 1981 come after May 09, 1991? **Predicted 4:** no ✓

Table 5: Error cases of the previous models on our dataset. It is noted that there are no existing models that can perform on all the three levels. The results in examples #3 and #4 are from the two models HGN and NumNet+.