Compositional Questions Do Not Necessitate Multi-hop Reasoning

Sewon Min*[1], Eric Wallace*[2], Sameer Singh[3],
Matt Gardner[2], Hannaneh Hajishirzi[1,2], Luke Zettlemoyer[1]
[1]University of Washington
[2]Allen Institute for Artificial Intelligence
[3]University of California, Irvine
sewon@cs.washington.edu, ericw@allenai.org

Abstract

Multi-hop reading comprehension (RC) questions are challenging because they require reading and reasoning over multiple paragraphs. We argue that it can be difficult to construct large multi-hop RC datasets. For example, even highly compositional questions can be answered with a single hop if they target specific entity types, or the facts needed to answer them are redundant. Our analysis is centered on HotpotQA, where we show that single-hop reasoning can solve much more of the dataset than previously thought. We introduce a single-hop BERT-based RC model that achieves 67 F1—comparable to state-of-the-art multi-hop models. We also design an evaluation setting where humans are not shown all of the necessary paragraphs for the intended multi-hop reasoning but can still answer over 80% of questions. Together with detailed error analysis, these results suggest there should be an increasing focus on the role of evidence in multi-hop reasoning and possibly even a shift towards information retrieval style evaluations with large and diverse evidence collections.

Question: What is the former name of the animal whose habitat the Réserve Naturelle Lomako Yokokala was established to protect?
Paragraph 5: The Lomako Forest Reserve is found in Democratic Republic of the Congo. It was established in 1991 especially to protect the habitat of the Bonobo apes.
Paragraph 1: The bonobo ("Pan paniscus"), formerly called the **pygmy chimpanzee** and less often, the dwarf or gracile chimpanzee, is an endangered great ape and one of the two species making up the genus "Pan".

Figure 1: A HotpotQA example designed to require reasoning across two paragraphs. Eight spurious additional paragraphs (not shown) are provided to increase the task difficulty. However, since only one of the ten paragraphs is about an animal, one can immediately locate the answer in *Paragraph 1* using one hop. The full example is provided in Appendix A.

1 Introduction

Multi-hop reading comprehension (RC) requires reading and aggregating information over multiple pieces of textual evidence (Welbl et al., 2017; Yang et al., 2018; Talmor and Berant, 2018). In this work, we argue that it can be difficult to construct large multi-hop RC datasets. This is because multi-hop reasoning is a characteristic of both the question and the provided evidence; even highly compositional questions can be answered with a single hop if they target specific entity types, or the facts needed to answer them are redundant. For example, the question in Figure 1 is compositional: a plausible solution is to find "What animal's habitat was the Réserve Naturelle Lomako Yokokala established to protect?", and then answer "What is the former name of that animal?". However, when considering the evidence paragraphs, the question is solvable in a single hop by finding the only paragraph that describes an animal.

Our analysis is centered on HotpotQA (Yang et al., 2018), a dataset of mostly compositional questions. In its RC setting, each question is paired with two gold paragraphs, which should be needed to answer the question, and eight distractor paragraphs, which provide irrelevant evidence or incorrect answers. We show that single-hop reasoning can solve much more of this dataset than previously thought. First, we design a single-hop QA model based on BERT (Devlin et al., 2018), which, despite having no ability to reason across paragraphs, achieves performance competitive with the state of the art. Next, we present an evaluation demonstrating that humans can solve over 80% of questions when we withhold one of the gold paragraphs.

To better understand these results, we present a detailed analysis of why single-hop reasoning works so well. We show that questions include redundant facts which can be ignored when com-

*Equal Contribution.

puting the answer, and that the fine-grained entity types present in the provided paragraphs in the RC setting often provide a strong signal for answering the question, e.g., there is only one animal in the given paragraphs in Figure 1, allowing one to immediately locate the answer using one hop.

This analysis shows that more carefully chosen distractor paragraphs would induce questions that require multi-hop reasoning. We thus explore an alternative method for collecting distractors based on adversarial paragraph selection. Although this appears to mitigate the problem, a single-hop model re-trained on these distractors can recover most of the original single-hop accuracy, indicating that these distractors are still insufficient. Another method is to consider very large distractor sets such as all of Wikipedia or the entire Web, as done in open-domain HotpotQA and ComplexWebQuestions (Talmor and Berant, 2018). However, this introduces additional computational challenges and/or the need for retrieval systems. Finding a small set of distractors that induce multi-hop reasoning remains an open challenge that is worthy of follow up work.

2 Related Work

Large-scale RC datasets (Hermann et al., 2015; Rajpurkar et al., 2016; Joshi et al., 2017) have enabled rapid advances in neural QA models (Seo et al., 2017; Xiong et al., 2018; Yu et al., 2018; Devlin et al., 2018). To foster research on reasoning across *multiple* pieces of text, multi-hop QA has been introduced (Kočiskỳ et al., 2018; Talmor and Berant, 2018; Yang et al., 2018). These datasets contain compositional or "complex" questions. We demonstrate that these questions do not necessitate multi-hop reasoning.

Existing multi-hop QA datasets are constructed using knowledge bases, e.g., WikiHop (Welbl et al., 2017) and ComplexWebQuestions (Talmor and Berant, 2018), or using crowd workers, e.g., HotpotQA (Yang et al., 2018). WikiHop questions are posed as triples of a relation and a head entity, and the task is to determine the tail entity of the relationship. ComplexWebQuestions consists of open-domain compositional questions, which are constructed by increasing the complexity of SPARQL queries from WebQuestions (Berant et al., 2013). We focus on HotpotQA, which consists of multi-hop questions written to require reasoning over two paragraphs from Wikipedia.

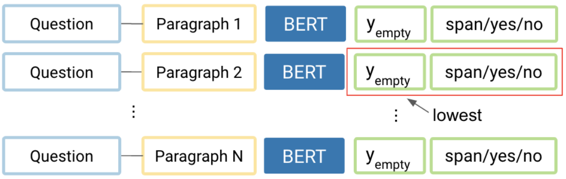

Figure 2: Our model, single-paragraph BERT, reads and scores each paragraph independently. The answer from the paragraph with the lowest y_{empty} score is chosen as the final answer.

Parallel research from Chen and Durrett (2019) presents similar findings on HotpotQA. Our work differs because we conduct human analysis to understand why questions are solvable using single-hop reasoning. Moreover, we show that selecting distractor paragraphs is difficult using current retrieval methods.

3 Single-paragraph QA

This section shows the performance of a single-hop model on HotpotQA.

3.1 Model Description

Our model, single-paragraph BERT, scores and answers each paragraph independently (Figure 2). We then select the answer from the paragraph with the best score, similar to Clark and Gardner (2018).[1]

The model receives a question $Q = [q_1, .., q_m]$ and a single paragraph $P = [p_1, ..., p_n]$ as input. Following Devlin et al. (2018), $S = [q_1, ..., q_m, [\texttt{SEP}], p_1, ..., p_n]$, where [SEP] is a special token, is fed into BERT:

$$S' = \text{BERT}(S) \in \mathbb{R}^{h \times (m+n+1)},$$

where h is the hidden dimension of BERT. Next, a classifier uses max-pooling and learned parameters $W_1 \in \mathbb{R}^{h \times 4}$ to generate four scalars:

$$[y_{\text{span}}; y_{\text{yes}}; y_{\text{no}}; y_{\text{empty}}] = W_1 \text{maxpool}(S'),$$

where $y_{\text{span}}, y_{\text{yes}}, y_{\text{no}}$ and y_{empty} indicate the answer is either a span, `yes`, `no`, or no answer. An extractive paragraph span, `span`, is obtained separately following Devlin et al. (2018). The final model outputs are a scalar value y_{empty} and a text of either `span`, `yes` or `no`, based on which of $y_{\text{span}}, y_{\text{yes}}, y_{\text{no}}$ has the largest value.

[1] Full details in Appendix B. Code available at `https://github.com/shmsw25/single-hop-rc`.

Model	Distractor F1	Open F1
Single-paragraph BERT*	67.08	38.40
BiDAF*	58.28	34.36
BiDAF	58.99	32.89
GRN	66.71	36.48
QFE	68.06	38.06
DFGN + BERT	68.49	-
MultiQA	-	40.23
DecompRC	69.63	40.65
BERT Plus	69.76	-
Cognitive Graph	-	48.87

Table 1: F1 scores on HOTPOTQA. * indicates the result is on the validation set; the other results are on the hidden test set shown in the official leaderboard.

For a particular HOTPOTQA example, we run single-paragraph BERT on each paragraph in parallel and select the answer from the paragraph with the smallest y_{empty}.

3.2 Model Results

HOTPOTQA has two settings: a distractor setting and an open-domain setting.

Distractor Setting The HOTPOTQA distractor setting pairs the two paragraphs the question was written for (*gold paragraphs*) with eight spurious paragraphs selected using TF-IDF similarity with the question (*distractors*). Our single-paragraph BERT model achieves 67.08 F1, comparable to the state-of-the-art (Table 1).[2] This indicates the majority of HOTPOTQA questions are answerable in the distractor setting using a single-hop model.

Open-domain Setting The HOTPOTQA open-domain setting (*Fullwiki*) does not provide a set of paragraphs—all of Wikipedia is considered. We follow Chen et al. (2017) and retrieve paragraphs using bigram TF-IDF similarity with the question.

We use the single-paragraph BERT model trained in the distractor setting. We also fine-tune the model using incorrect paragraphs selected by the retrieval system. In particular, we retrieve 30 paragraphs and select the eight paragraphs with the lowest y_{empty} scores predicted by the trained model. Single-paragraph BERT achieves 38.06 F1 in the open-domain setting (Table 1). This shows that the open-domain setting is challenging for our single-hop model and is worthy of future study.

4 Compositional Questions Are Not Always Multi-hop

This section provides a human analysis of HOTPOTQA to understand what phenomena enable single-hop answer solutions. HOTPOTQA contains two question types, *Bridge* and *Comparison*, which we evaluate separately.

4.1 Categorizing Bridge Questions

Bridge questions consist of two paragraphs linked by an entity (Yang et al., 2018), e.g., Figure 1. We first investigate single-hop human performance on HOTPOTQA bridge questions using a human study consisting of NLP graduate students. Humans see the paragraph that contains the answer span and the eight distractor paragraphs, but do not see the other gold paragraph. As a baseline, we show a different set of people the same questions in their standard ten paragraph form.

On a sample of 200 bridge questions from the validation set, human accuracy shows marginal degradation when using only one hop: humans obtain **87.37 F1** using all ten paragraphs and **82.06 F1** when using only nine (where they only see a single gold paragraph). This indicates humans, just like models, are capable of solving bridge questions using only one hop.

Next, we manually categorize what enables single-hop answers for 100 bridge validation examples (taking into account the distractor paragraphs), and place questions into four categories (Table 2).

Multi-hop 27% of questions require multi-hop reasoning. The first example of Table 2 requires locating the university where "Ralph Hefferline" was a psychology professor, and multiple universities are provided as distractors. Therefore, the answer cannot be determined in one hop.[3]

Weak Distractors 35% of questions allow single-hop answers in the distractor setting, mostly by entity type matching. Consider the question in the second row of Table 2: in the ten provided paragraphs, only one actress has a government position. Thus, the question is answerable without considering the film "Kiss and Tell." These examples may become multi-hop in the open-domain setting, e.g., there are numerous actresses with a government position on Wikipedia.

[2] Results as of March 4th, 2019.

[3] It is possible that a single-hop model can do well by randomly guessing between two or three well-typed options, but we do not evaluate that strategy here.

Type	Question	%
Multi-hop	Ralph Hefferline was a psychology professor at a university that is located in what city?	27
Weak distractors	What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?	35
Redundant evidence	Kaiser Ventures corporation was founded by an American industrialist who became known as the father of modern American shipbuilding?	26
Non-compositional 1-hop	When was Poison's album 'Shut Up, Make Love' released?	8

Table 2: We categorize bridge questions while taking the paragraphs into account. We exclude 4% of questions that we found to have incorrect or ambiguous answer annotations. See Section 4.1 for details on question types.

Type	Question	%	F1
Multi-hop	Who was born first, Arthur Conan Doyle or Penelope Lively?	45	54.46
Context-dependent	Are Hot Rod and the Memory of Our People both magazines?	36	56.16
Single-hop	Which writer was from England, Henry Roth or Robert Erskine Childers?	17	70.54

Table 3: We automatically categorize comparison questions using rules (2% cannot be automatically categorized). Single-paragraph BERT achieves near chance accuracy on multi-hop questions but exploits single-hop ones.

To further investigate entity type matching, we reduce the question to the first five tokens starting from the wh-word, following Sugawara et al. (2018). Although most of these reduced questions appear void of critical information, the F1 score of single-paragraph BERT only degrades about 15 F1 from 67.08 to 52.13.

Redundant Evidence 26% of questions are compositional but are solvable using only part of the question. For instance, in the third example of Table 2 there is only a single founder of "Kaiser Ventures." Thus, one can ignore the condition on "American industrialist" and "father of modern American shipbuilding." This category differs from the weak distractors category because its questions are single-hop regardless of the distractors.

Non-compositional Single-hop 8% of questions are non-compositional and single-hop. In the last example of Table 2, one sentence contains all of the information needed to answer correctly.

4.2 Categorizing Comparison Questions

Comparison questions require quantitative or logical comparisons between two quantities or events. We create rules (Appendix C) to group comparison questions into three categories: questions which require multi-hop reasoning (*multi-hop*), may require multi-hop reasoning (*context-dependent*), and require single-hop reasoning (*single-hop*).

Many comparison questions are multi-hop or context-dependent multi-hop, and single-paragraph BERT achieves near chance accuracy on these types of questions (Table 3).[4] This shows that most comparison questions are not solvable by our single-hop model.

Evaluation Data	Training Data	
	Original	Adversarial
Original	67.08	59.12
Adversarial	46.84	60.10
+ Type	40.73	58.42

Table 4: We train on HOTPOTQA using standard distractors (*Original*) or using adversarial distractors (*Adversarial*). The model is then tested on the original distractors, adversarial distractors, or adversarial distractors with filtering by entity type (+ *Type*).

5 Can We Find Better Distractors?

In Section 4.1, we identify that 35% of bridge examples are solvable using single-hop reasoning due to weak distractor paragraphs. Here, we attempt to automatically correct these examples by choosing new distractor paragraphs which are likely to trick our single-paragraph model.

Adversarial Distractors We select the top-50 first paragraphs of Wikipedia pages using TF-IDF similarity with the question, following the original HOTPOTQA setup. Next, we use single-paragraph BERT to adversarially select the eight distractor paragraphs from these 50 candidates. In particular, we feed each paragraph to the model and select

[4] Comparison questions test mainly binary relationships.

the paragraphs with the lowest y_{empty} score (i.e., the paragraphs that the model thinks contain the answer). These paragraphs are dissimilar to the original distractors—there is a 9.82% overlap.

We report the F1 score of single-paragraph BERT on these new distractors in Table 4: the accuracy declines from 67.08 F1 to 46.84 F1. However, when the same procedure is done on the training set and the model is re-trained, the accuracy increases to 60.10 F1 on the adversarial distractors.

Type Distractors We also experiment with filtering the initial list of 50 paragraph to ones whose entity type (e.g., person) matches that of the gold paragraphs. This can help to eliminate the entity type bias described in Section 4.1. As shown in Table 4, the original model's accuracy degrades significantly (drops to 40.73 F1). However, similar to the previous setup, the model trained on the adversarially selected distractors can recover most of its original accuracy (increases to 58.42 F1).

These results show that single-paragraph BERT can struggle when the distribution of the distractors changes (e.g., using adversarial selection rather than only TF-IDF). Moreover, the model can somewhat recover its original accuracy when re-trained on distractors from the new distribution.

6 Conclusions

In summary, we demonstrate that question compositionality is not a sufficient condition for multi-hop reasoning. Instead, future datasets must carefully consider what evidence they provide in order to ensure multi-hop reasoning is required. There are at least two different ways to achieve this.

Open-domain Questions Our single-hop model struggles in the open-domain setting. We largely attribute this to the insufficiencies of standard TF-IDF retrieval for multi-hop questions. For example, we fail to retrieve the paragraph about "Bonobo apes" in Figure 1, because the question does not contain terms about "Bonobo apes." Table 5 shows that the model achieves 39.12 F1 given 500 retrieved paragraphs, but achieves 53.12 F1 when additional two gold paragraphs are given, demonstrating the significant effect of failure to retrieve gold paragraphs. In this context, we suggest that future work can explore better retrieval methods for multi-hop questions.

Setting	**F1**
Distractor	67.08
Open-domain 10 Paragraphs	38.40
Open-domain 500 Paragraphs	39.12
+ Gold Paragraph	53.12

Table 5: The accuracy of single-paragraph BERT in different open-domain retrieval settings. TF-IDF often fails to retrieve the gold paragraphs even when using 500 candidates.

Retrieving Strong Distractors Another way to ensure multi-hop reasoning is to select strong distractor paragraphs. For example, we found 35% of bridge questions are currently single-hop but may become multi-hop when combined with stronger distractors (Section 4.1). However, as we demonstrate in Section 5, selecting strong distractors for RC questions is non-trivial. We suspect this is also due to the insufficiencies of standard TF-IDF retrieval for multi-hop questions. In particular, Table 5 shows that single-paragraph BERT achieves 53.12 F1 even when using 500 distractors (rather than eight), indicating that 500 distractors are still insufficient. In this end, future multi-hop RC datasets can develop improved methods for distractor collection.

Acknowledgements

This research was supported by ONR (N00014-18-1-2826, N00014-17-S-B001), NSF (IIS-1616112, IIS-1252835, IIS-1562364), ARO (W911NF-16-1-0121), an Allen Distinguished Investigator Award, Samsung GRO and gifts from Allen Institute for AI, Google, and Amazon.

The authors would like to thank Shi Feng, Nikhil Kandpal, Victor Zhong, the members of AllenNLP and UW NLP, and the anonymous reviewers for their valuable feedback.

References

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *EMNLP*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *ACL*.

Jifan Chen and Greg Durrett. 2019. Understanding dataset design choices for multi-hop reasoning. In *NAACL*.

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In *ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *TACL*.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *EMNLP*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *ICLR*.

Saku Sugawara, Kentaro fInui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *EMNLP*.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *NAACL*.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2017. Constructing Datasets for Multi-hop Reading Comprehension Across Documents. In *TACL*.

Caiming Xiong, Victor Zhong, and Richard Socher. 2018. DCN+: Mixed objective and deep residual coattention for question answering. In *ICLR*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. In *ICLR*.

A Example Distractor Question

We present the full example from Figure 1 below. Paragraphs 1 and 5 are the two gold paragraphs.

Question What is the former name of the animal whose habitat the Réserve Naturelle Lomako Yokokala was established to protect?

Answer pygmy chimpanzee

(Gold Paragraph) Paragraph 1 The bonobo (or ; "Pan paniscus"), formerly called the **pygmy chimpanzee** and less often, the dwarf or gracile chimpanzee, is an endangered great ape and one of the two species making up the genus "Pan"; the other is "Pan troglodytes", or the common chimpanzee. Although the name "chimpanzee" is sometimes used to refer to both species together, it is usually understood as referring to the common chimpanzee, whereas "Pan paniscus" is usually referred to as the bonobo.

Paragraph 2 The Carriére des Nerviens Regional Nature Reserve (in French "Réserve naturelle régionale de la carriére des Nerviens") is a protected area in the Nord-Pas-de-Calais region of northern France. It was established on 25 May 2009 to protect a site containing rare plants and covers just over 3 ha. It is located in the municipalities of Bavay and Saint-Waast in the Nord department.

Paragraph 3 Céreste (Occitan: "Ceirésta") is a commune in the Alpes-de-Haute-Provence department in southeastern France. It is known for its rich fossil beds in fine layers of "Calcaire de Campagne Calavon" limestone, which are now protected by the Parc naturel régional du Luberon and the Réserve naturelle géologique du Luberon.

Paragraph 4 The Grand Cote National Wildlife Refuge (French: "Réserve Naturelle Faunique Nationale du Grand- Cote") was established in 1989 as part of the North American Waterfowl Management Plan. It is a 6000 acre reserve located in Avoyelles Parish, near Marksville, Louisiana, in the United States.

(Gold Paragraph) Paragraph 5 The Lomako Forest Reserve is found in Democratic Republic of the Congo. It was established in 1991 especially to protect the habitat of the Bonobo apes. This site covers 3,601.88 km^2.

Paragraph 6 Guadeloupe National Park (French: "Parc national de la Guadeloupe") is a national park in Guadeloupe, an overseas department of France located in the Leeward Islands of the eastern Caribbean region. The Grand Cul-de-Sac Marin Nature Reserve (French: "Réserve Naturelle du Grand Cul-de-Sac Marin") is a marine protected area adjacent to the park and administered in conjunction with it. Together, these protected areas comprise the Guadeloupe Archipelago (French: "l'Archipel de la Guadeloupe") biosphere reserve.

Paragraph 7 La Désirade National Nature Reserve (French: "Réserve naturelle nationale de La Désirade") is a reserve in Désirade Island in Guadeloupe. Established under the Ministerial Decree No. 2011-853 of 19 July 2011 for its special geological features it has an area of 62 ha. The reserve represents the geological heritage of the Caribbean tectonic plate, with a wide spectrum of rock formations, the outcrops of volcanic activity being remnants of the sea level oscillations. It is one of thirty three geosites of Guadeloupe.

Paragraph 8 La Tortue ou l'Ecalle or Ile Tortue is a small rocky islet off the northeastern coast of Saint Barthélemy in the Caribbean. Its highest point is 35 m above sea level. Referencing tortoises, it forms part of the Réserve naturelle nationale de Saint-Barthélemy with several of the other northern islets of St Barts.

Paragraph 9 Nature Reserve of Saint Bartholomew (Réserve Naturelle de Saint-Barthélemy) is a nature reserve of Saint Barthélemy (RNN 132), French West Indies, an overseas collectivity of France.

Paragraph 10 Ile Fourchue, also known as Ile Fourche is an island between Saint-Barthélemy and Saint Martin, belonging to the Collectivity of Saint Barthélemy. The island is privately owned. The only inhabitants are some goats. The highest point is 103 meter above sea level. It is situated within Réserve naturelle nationale de Saint-Barthélemy.

B Full Model Details

Single-paragraph BERT is a pipeline which first retrieves a single paragraph using a classifier and then selects the associated answer. Formally, the model receives a question $Q = [q_1, .., q_m]$ and a single paragraph $P = [p_1, ..., p_n]$ as input. The question and paragraph are merged into a single

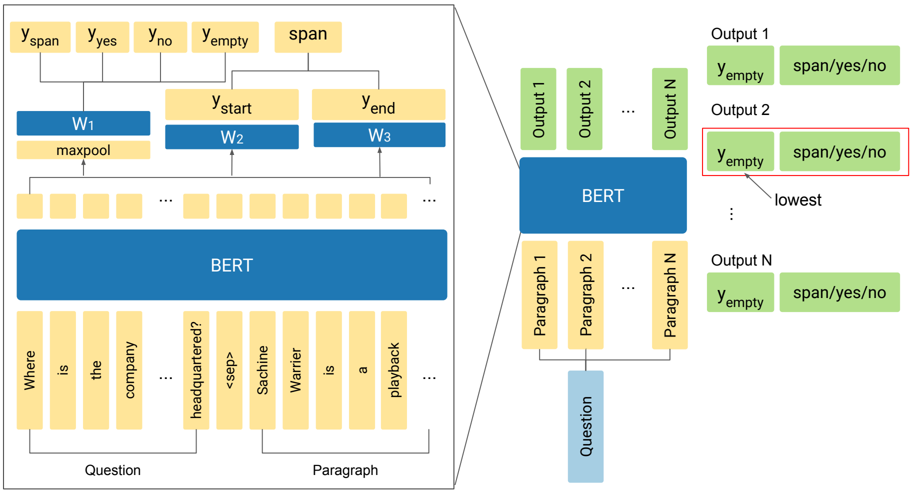

Figure 3: Single-paragraph BERT reads and scores each paragraph independently. The answer from the paragraph with the lowest y^{empty} score is chosen as the final answer.

sequence, $S = [q_1, ..., q_m, \text{[SEP]}, p_1, ..., p_n]$, where [SEP] is a special token indicating the boundary. The sequence is fed into BERT-BASE:

$$S' = \text{BERT}(S) \in \mathbb{R}^{h \times (m+n+1)},$$

where h is the hidden dimension of BERT. Next, a classifier uses max-pooling and learned parameters $W_1 \in \mathbb{R}^{h \times 4}$ to generate four scalars:

$$[y_{\text{span}}; y_{\text{yes}}; y_{\text{no}}; y_{\text{empty}}] = W_1 \text{maxpool}(S'),$$

where $y_{\text{span}}, y_{\text{yes}}, y_{\text{no}}$ and y_{empty} indicate the answer is either a span, `yes`, `no`, or no answer.

A candidate answer span is then computed separately from the classifier. We define

$$p_{\text{start}} = \text{Softmax}(W_2 S')$$
$$p_{\text{end}} = \text{Softmax}(W_3 S'),$$

where $W_2, W_3 \in \mathbb{R}^h$ are learned parameters. Then, y_{start} and y_{end} are obtained:

$$y_{\text{start}}, y_{\text{end}} = \underset{i \leq j}{\arg\max}\, p^i_{\text{start}} p^j_{\text{end}}$$

where p^i_{start} and p^j_{end} indicate the i-th element of p_{start} and j-th element of p_{end}, respectively.

We now have four scalar values y_{span}, y_{yes}, y_{no}, and y_{empty} and a span from the paragraph span $= [S_{y_{\text{start}}}, \ldots, S_{y_{\text{end}}}]$.

For HOTPOTQA, the input is a question and N context paragraphs. We create a batch of size N, where each entry is a question and a single paragraph. Denote the ouput from i-th entry as $y^i_{\text{span}}, y^i_{\text{yes}}, y^i_{\text{no}}, y^i_{\text{empty}}$ and span^i. The final answer is selected as:

$$\begin{aligned} j &= \text{argmin}_i(y^i_{\text{empty}}) \\ y_{\max} &= \max(y^j_{\text{span}}, y^j_{\text{yes}}, y^j_{\text{no}}) \\ \text{answer} &= \begin{cases} \text{span}^j & \text{if} y^j_{\text{span}} = y_{\max} \\ \texttt{yes} & \text{if} y^j_{\text{yes}} = y_{\max} \\ \texttt{no} & \text{if} y^j_{\text{no}} = y_{\max} \end{cases} \end{aligned}$$

During training, y^i_{empty} is set to 0 for the paragraph which contains the answer span and 1 otherwise.

Implementation Details We use PyTorch (Paszke et al., 2017) based on Hugging Face's implementation.[5] We use Adam (Kingma and Ba, 2015) with learning rate 5×10^{-5}. We lowercase the input and set the maximum sequence length $|S|$ to 300. If a sequence is longer than 300, we split it into multiple sequences and treat them as different examples.

C Categorizing Comparison Questions

This section describes how we categorize comparison questions. We first identify ten question operations that sufficiently cover comparison questions (Table 6). Next, for each question, we extract the two entities under comparison using the Spacy[6] NER tagger on the question and the two HOTPOTQA supporting facts. Using these extracted

[5] https://github.com/huggingface/pytorch-pretrained-BERT

[6] https://spacy.io/

Operation & Example
Numerical Questions Operations: `Is greater / Is smaller / Which is greater / Which is smaller` Example (`Which is smaller`): Who was born first, Arthur Conan Doyle or Penelope Lively?
Logical Questions Operations: `And / Or / Which is true` Example (`And`): Are Hot Rod and the Memory of Our People both magazines?
String Questions Operations: `Is equal / Not equal / Intersection` Example (`Is equal`): Are Cardinal Health and Kansas City Southern located in the same state?

Table 6: The question operations used for categorizing comparison questions.

Algorithm 1 Algorithm for Identifying Question Operations

```
procedure CATEGORIZE(question, entity1, entity2)
    coordination, preconjunct ← $f$(question, entity1, entity2)
    Determine if the question is either question or both question from coordination and preconjunct
    head entity ← $f_{head}$(question, entity1, entity2)
    if more, most, later, last, latest, longer, larger, younger, newer, taller, higher in question then
        if head entity exists then discrete_operation ← Which is greater
        else discrete_operation ← Is greater
    else if less, earlier, earliest, first, shorter, smaller, older, closer in question then
        if head entity exists then discrete_operation ← Which is smaller
        else discrete_operation ← Is smaller
    else if head entity exists then
        discrete_operation ← Which is true
    else if question is not yes/no question and asks for the property in common then
        discrete_operation ← Intersection
    else if question is yes/no question then
        Determine if question asks for logical comparison or string comparison
        if question asks for logical comparison then
            if either question then discrete_operation ← Or
            else if both question then discrete_operation ← And
        else if question asks for string comparison then
            if asks for same? then discrete_operation ← Is equal
            else if asks for difference? then discrete_operation ← Not equal
    return discrete_operation
```

entities, we identity the suitable question operation following Algorithm 1.

Based on the identified operation, questions are classified into multi-hop, context-dependent multi-hop, or single-hop. First, numerical questions are always multi-hop (e.g., first example of Table 6). Next, the operations `And, Or, Is equal,` and `Not equal` are context-dependent multi-hop. For instance, in the second example of Table 6, if "Hot Rod" is not a magazine, one can immediately answer `No`. Finally, the operations `Which is true` and `Intersection` are single-hop because they can be answered using one paragraph regardless of the context. For instance, in the third example of Table 6, if Henry Roth's paragraph explains he is from England, one can answer Henry Roth, otherwise, the answer is Robert Erskine Childers.