Reasoning Over Paragraph Effects in Situations

Kevin Lin, Oyvind Tafjord, Peter Clark, and **Matt Gardner**
Allen Institute for Artificial Intelligence
{kevinl,oyvindt,peterc,mattg}@allenai.org

Abstract

A key component of successfully reading a passage of text is the ability to apply knowledge gained from the passage to a new situation. In order to facilitate progress on this kind of reading, we present **ROPES**, a challenging benchmark for reading comprehension targeting **R**easoning **O**ver **P**aragraph **E**ffects in **S**ituations. We target expository language describing causes and effects (e.g., "animal pollinators increase efficiency of fertilization in flowers"), as they have clear implications for new situations. A system is presented a background passage containing at least one of these relations, a novel situation that uses this background, and questions that require reasoning about effects of the relationships in the background passage in the context of the situation. We collect background passages from science textbooks and Wikipedia that contain such phenomena, and ask crowd workers to author situations, questions, and answers, resulting in a 14,322 question dataset. We analyze the challenges of this task and evaluate the performance of state-of-the-art reading comprehension models. The best model performs only slightly better than randomly guessing an answer of the correct type, at 61.6% F1, well below the human performance of 89.0%.

Background: Scientists think that the earliest flowers attracted **insects and other animals, which spread pollen** from flower to flower. **This greatly increased the efficiency of fertilization over wind-spread pollen**, which might or might not actually land on another flower. **To take better advantage of this animal labor, plants evolved traits such as brightly colored petals to attract pollinators**. In exchange for pollination, flowers gave the pollinators nectar.

Situation: Last week, John visited the national park near his city. He saw many flowers. His guide explained him that there are two categories of flowers, category A and category B. **Category A flowers spread pollen via wind, and category B flowers spread pollen via animals**.

Question: Would category B flower have **more or less efficient fertilization** than category A flower?
Answer: more

Question: Would category A flower have **more or less efficient fertilization** than category B flower?
Answer: less

Question: Which category of flowers would be more likely to have **brightly colored petals**?
Answer: Category B

Question: Which category of flowers would be less likely to have **brightly colored petals**?
Answer: Category A

Figure 1: Example questions in **ROPES**.

1 Introduction

Comprehending a passage of text requires being able to understand the implications of the passage on other text that is read. For example, after reading a background passage about how animal pollinators increase the efficiency of fertilization in flowers, a human can easily deduce that given two types of flowers, one that attracts animal pollinators and one that does not, the former is likely to have a higher efficiency in fertilization (Figure 1). This kind of reasoning however, is still challenging for state-of-the-art reading comprehension models.

Recent work in reading comprehension has seen impressive results, with models reaching human performance on well-established datasets (Devlin et al., 2019; Wang et al., 2017; Chen et al., 2016), but so far has mostly focused on extracting local predicate-argument structure, without the need to apply what was read to outside context.

We introduce **ROPES**[1], a reading comprehension challenge that focuses on understanding causes and effects in an expository paragraph, requiring systems to apply this understanding to

[1] https://allennlp.org/ropes

novel situations. If a new situation describes an occurrence of the cause, then the system should be able to reason over the effects if it has properly understood the background passage.

We constructed **ROPES** by first collecting background passages from science textbooks and Wikipedia articles that describe causal relationships. We showed these paragraphs to crowd workers and asked them to write situations that involve the relationships found in the background passage, and questions that connect the situation and the background using the causal relationships. The answers are spans from either the situation or the question. The dataset consists of 14,322 questions from various domains, mostly in science and economics.

In analyzing the data, we find (1) that there are a variety of cause / effect relationship types described; (2) that there is a wide range of difficulties in matching the descriptions of these phenomena between the background, situation, and question; and (3) that there are several distinct kinds of reasoning over causes and effects that appear.

To establish baseline performance on this dataset, we use a reading comprehension model based on RoBERTa (Liu et al., 2019), reaching an accuracy of 61.6% F_1. Most questions are designed to have two sensible answer choices (eg. "more" vs. "less"), so this performance is little better than randomly picking one of the choices. Expert humans achieved an average of 89.0% F_1 on a random sample.

2 Related Work

Reading comprehension There are many reading comprehension datasets (Richardson et al., 2013; Rajpurkar et al., 2016; Kwiatkowski et al., 2019; Dua et al., 2019), the majority of which principally require understanding local predicate-argument structure in a passage of text. The success of recent models suggests that machines are becoming capable of this level of understanding. **ROPES** challenges reading comprehension models to handle more difficult phenomena: understanding the *implications* of a passage of text. **ROPES** is also particularly related to datasets focusing on "multi-hop reasoning" (Yang et al., 2018; Khashabi et al., 2018), as by construction answering questions in **ROPES** requires connecting information from multiple parts of a given passage.

The most closely related datasets to **ROPES** are ShARC (Saeidi et al., 2018), OpenBookQA (Mihaylov et al., 2018), and QuaRel (Tafjord et al., 2019). ShARC shares the same goal of understanding causes and effects (in terms of specified rules), but frames it as a dialogue where the system has to also generate questions to gain complete information. OpenBookQA, similar to **ROPES**, requires reading scientific facts, but it is focused on a *retrieval* problem where a system must find the right fact for a question (and some additional common sense fact), whereas **ROPES** targets *reading* a given, complex passage of text, with no retrieval involved. QuaRel is also focused on reasoning about situational effects in a question-answering setting, but the "causes" are all pre-specified, not read from a background passage, so the setting is limited.

Recognizing textual entailment The application of causes and effects to new situations has a strong connection to notions of entailment—**ROPES** tries to get systems to understand what is entailed by an expository paragraph. The setup is fundamentally different, however: instead of giving systems pairs of sentences to classify as entailed or not, as in the traditional formulation (Dagan et al., 2006; Bowman et al., 2015, *inter alia*), we give systems questions whose answers require understanding the entailment.

3 Data Collection

Background passages: We automatically scraped passages from science textbooks[2] and Wikipedia that contained causal connectives eg. "causes," "leads to," and keywords that signal qualitative relations, e.g. "increases," "decreases."[3]. We then manually filtered out the passages that do not have at least one relation. The passages can be categorized into physical science (49%), life science (45%), economics (5%) and other (1%). In total, we collected over 1,000 background passages.

Crowdsourcing questions We used Amazon Mechanical Turk (AMT) to generate the situations, questions, and answers. The AMT workers were given background passages and asked to write situations that involved the relation(s) in the background passage. The AMT workers then authored questions about the situation that required both the

[2] We used life science and physical science concepts from www.ck12.org, and biology, chemistry, physics, earth science, anatomy and physiology textbooks from openstax.org

[3] We scraped Wikipedia online in March and April 2019

Statistic	Train	Dev	Test
# of annotators	7	2	2
# of situations	1411	203	300
# of questions	10924	1688	1710
avg. background length	121.6	90.7	123.1
avg. situation length	49.1	63.4	55.6
avg. question length	10.9	12.4	10.6
avg. answer length	1.3	1.4	1.4
background vocabulary size	8616	2008	3988
situation vocabulary size	6949	1077	2736
question vocabulary size	1457	1411	1885

Table 1: Key statistics of **ROPES**. In total there were 588 background passages selected by the workers.

Type	Background
C (70%)	Scientists think that the earliest flowers **attracted insects and other animals**, which spread pollen from flower to flower. This greatly **increased the efficiency of fertilization** over wind-spread pollen ...
Q (4%)	... As **decibel levels get higher**, **sound waves have greater intensity** and **sounds are louder**. ...
C&Q (26%)	... Predators can be keystone species . These are species that can have a large effect on the balance of organisms in an ecosystem. For example, if all of **the wolves are removed from a population**, then **the population of deer or rabbits may increase**...

Table 2: Types of relations in the background passages. **C** refers to causal relations and **Q** refers to qualitative relations.

background and the situation to answer. In each human intelligence task (HIT), AMT workers are given 5 background passages to select from and are asked to create a total of 10 questions. To mitigate the potential for easy lexical shortcuts in the dataset, the workers were encouraged via instructions to write questions in *minimal pairs*, where a very small change in the question results in a different answer. Two examples of these pairs are given in Figure 1: switching “more” to “less” results in the opposite flower being the correct answer to the question.

4 Dataset Analysis

We qualitatively and quantitatively analyze the phenomena that occur in **ROPES**. Table 1 shows the key statistics of the dataset. We randomly sample 100 questions and analyze the type of relation in the background, grounding in the situation, and reasoning required to answer the question.

Type	Background	Situation
Explicit (67%)	As **decibel levels get higher**, sound waves have greater intensity and sounds are louder.	...First, he went to stage one, where the music was playing in **high decibel**.
Common sense (13%)	... if we want to convert a substance from a gas to a liquid or from a **liquid to a solid**, we remove energy from the system	... She remembered they would be needing ice so she **grabbed and empty ice tray and filled it**...
Lexical gap (20%)	... **Continued exercise** is necessary to maintain bigger, stronger muscles...	... Mathew goes to the gym ... does **very intensive workouts**.

Table 3: Types of grounding found in **ROPES**.

Background passages We manually annotate whether the relation in the background passage being asked about is causal (a clear cause and effect in the background), qualitative (e.g., as X increases, Y decreases), or both. Table 2 shows the breakdown of the kinds of relations in the dataset.

Grounding To successfully apply the relation in the background to a situation, the system needs to be able to ground the relation to parts of the situation. To do this, the model has to either find an *explicit* mention of the cause/effect from the background and associate it with some property, use a *common sense fact*, or overcome a large *lexical gap* to connect them. Table 3 shows examples and breakdown of these three phenomena.

Question reasoning Table 4 shows the breakdown and examples of the main types of questions by the types of reasoning required to answer them. In an *effect comparison*, two entities are each associated with an occurrence or absence of the cause described in the background and the question asks to compare the effects on the two entities. Similarly, in a *cause comparison*, two entities are each associated with an occurrence or absence of the effect described in the background and the question compares the causes of the occurrence or absence. In an *effect prediction*, the question asks to directly predict the effect on an occurrence of the cause on an entity in the situation. Finally, in *cause prediction*, the question asks to predict the cause of an occurrence of the effect on an entity in the situation. The majority of the examples are effect or cause comparison questions; these are challenging, as they require the model to ground two occurrences of causes or effects.

Reasoning	**Background**	**Situation**	**Question**	**Answer**
Effect comparison (71%)	... gas atoms change to ions that can carry an electric current. The current causes the Geiger counter to click. **The faster the clicks occur**, the **higher the level of radiation.**	... Location A had **very high radiation**; location B had low radiation	Would location A have **faster** or slower clicks than location B?	faster
Effect prediction (5%)	... **Continued exercise** is necessary to maintain **bigger, stronger muscles.** ...	... Mathew goes to the gym 5 times a week and **does very intensive workouts.** Damen on the other hand does not go to the gym at all and lives a mostly sedentary lifestyle.	Given Mathew suffers an injury while working out and **cannot go to the gym for 3 months**, will Mathews strength increase or **decrease**?	decrease
Cause comparison (15%)	... This **carbon dioxide is then absorbed by the oceans**, which **lowers the pH of the water**...	The biologists found out that the Indian Ocean had a **lower water pH** than it did a decade ago, and it became acidic. The water in the Arctic ocean still had a **neutral to basic pH**.	Which ocean has a **lower content of carbon dioxide** in its waters?	Arctic
Cause prediction (1%)	... Conversely, if we want to convert a substance from a gas to a liquid or from a **liquid to a solid**, we **remove energy from the system** and decrease the temperature. ...	... she grabbed and empty ice tray and filled it. As she walked over to the freezer ... When she checked the tray later that day the **ice was ready**.	Did the freezer add or **remove** energy from the water?	remove
Other (8%)	... **Charging an object by touching it** with another charged object is called charging by **conduction**. ... **induction** allows a change in charge **without actually touching the charged and uncharged objects** to each other.	... In case A he used **conduction**, and in case B he used **induction**. In both cases he used same two objects. Finally, John tried to **charge his phone remotely**. He called this test as **case C**.	Which experiment would be less appropriate for **case C**, **case A** or **case B**?	case A

Table 4: Example questions and answers from **ROPES**, showing the relevant parts of the associated passage and the reasoning required to answer the question. In the last example, the situation grounds the desired outcome and asks which of two cases would achieve the desired outcome.

Dataset split In initial experiments, we found splitting the dataset based on the situations resulted in high scores due to annotator bias from prolific workers generating many examples (Geva et al., 2019). We follow their proposal and separate training set annotators from test set annotators, and find that models have difficulty generalizing to new workers.

5 Baseline performance

We use the RoBERTa question answering model proposed by Liu et al. (2019) as our baseline and concatenate the background and situation to form the passage, following their setup for SQuAD. To estimate the presence of annotation artifacts in our dataset (and as a potentially interesting future task where background reading is done up front), we also run the baseline without the background passage. Table 5 presents the results for the baselines, which are significantly lower than human performance. We also experiment with first fine-tuning on RACE (Lai et al., 2017) before fine-tuning on **ROPES**.

	Development		**Test**	
	EM	**F1**	**EM**	**F1**
RoBERTa BASE	38.0	53.5	35.8	45.5
- background	40.7	59.3	33.7	46.1
RoBERTa LARGE	59.7	70.2	55.4	61.1
- background	48.7	55.2	53.6	60.4
+ RACE	60.1	73.5	55.5	61.6
Human	-	-	82.7	89.0

Table 5: Performance of baselines and human performance on the dev and test set.

Human performance is estimated by expert human annotation on 400 random questions with the same metrics as the baselines. None of the questions share the sample background or situation to ensure that the humans do not have an unfair advantage over the model by using knowledge of how the dataset is constructed, e.g. the fact that pairs of questions like in Table 1 will have opposite answers.

6 Conclusion

We present **ROPES**, a new reading comprehension benchmark containing 14,322 questions, which aims to test the ability of systems to apply knowledge from reading text in a new setting. We hope that **ROPES** will aide efforts in tying language and reasoning together for more comprehensive understanding of text.

7 Acknowledgements

We thank the anonymous reviewers for their feedback. We also thank Dheeru Dua and Nelson Liu for their assistance with the crowdsourcing setup, and Kaj Bostrom for the human evaluation. We are grateful for discussion with other AllenNLP and Aristo team members and the infrastructure provided by the Beaker team. Computations on beaker.org were supported in part by credits from Google Cloud.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

Danqi Chen, Jason Bolton, and Christopher D Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. *arXiv preprint arXiv:1606.02858*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. *Lecture Notes in Computer Science*, pages 177–190.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

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 *NAACL*.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *EMNLP*.

Daniel Khashabi, Snigdha Chaturvedi, Michael A. Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *NAACL-HLT*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. In *TACL*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

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

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.

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

Matthew Richardson, Christopher J. C. Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *EMNLP*.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. *arXiv preprint arXiv:1809.01494*.

Oyvind Tafjord, Peter Clark, Matt Gardner, Wen-tau Yih, and Ashish Sabharwal. 2019. Quarel: A dataset and models for answering questions about qualitative relationships. In *AAAI*.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198.

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