QUARTZ: An Open-Domain Dataset of Qualitative Relationship Questions

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

Abstract

We introduce the first open-domain dataset, called QUARTZ, for reasoning about textual qualitative relationships. QUARTZ contains general qualitative statements, e.g., “A sunscreen with a higher SPF protects the skin longer.”, twinned with 3864 crowdsourced situated questions, e.g., “Billy is wearing sunscreen with a lower SPF than Lucy. Who will be best protected from the sun?”, plus annotations of the properties being compared. Unlike previous datasets, the general knowledge is textual and not tied to a fixed set of relationships, and tests a system’s ability to comprehend and apply textual qualitative knowledge in a novel setting. We find state-of-the-art results are substantially (20%) below human performance, presenting an open challenge to the NLP community.

1 Introduction

Understanding and applying qualitative knowledge is a fundamental facet of intelligence. For example, we may read that exercise improves health, and thus decide to spend more time at the gym; or that larger cars cause more pollution, and thus decide to buy a smaller car to be environmentally sensitive. These skills require understanding the underlying qualitative relationships, and being able to apply them in specific contexts.

To promote research in this direction, we present the first *open-domain* dataset of qualitative relationship questions, called QUARTZ (“Qualitative Relationship Test set”)[1]. Unlike earlier work in qualitative reasoning, e.g., (Tafjord et al., 2019), the dataset is not restricted to a small, fixed set of relationships. Each question Q_i (2-way multiple choice) is grounded in a particular situation, and is paired with a sentence K_i expressing the general qualitative knowledge needed to answer it.

[1] Available at http://data.allenai.org/quartz/

Q: If Mona lives in a city that begins producing a greater amount of pollutants, what happens to the air quality in that city? (A) increases (B) decreases **[correct]**

K: More pollutants mean poorer air quality.

Annotations:

Q: [**MORE**, ”greater”, ”amount of pollutants”]
→ (A) [**MORE**, ”increases”, ”air quality”]
(B) [**LESS**, ”decreases”, ”air quality”]

K: [**MORE**, ”more”, ”pollutants”]
↔ [**LESS**, ”poorer”, ”air quality”]

Figure 1: QUARTZ contains situated qualitative questions, each paired with a gold background knowledge sentence and qualitative annotations.

Q_i and K_i are also annotated with the properties being compared (Figure 1). The property annotations serve as supervision for a potential semantic parsing based approach. The overall task is to answer the Q_i given the corpus $K = \{K_i\}$.

We test several state-of-the-art (BERT-based) models and find that they are still substantially (20%) below human performance. Our contributions are thus (1) the dataset, containing 3864 richly annotated questions plus a background corpus of 400 qualitative knowledge sentences; and (2) an analysis of the dataset, performance of BERT-based models, and a catalog of the challenges it poses, pointing the way towards solutions.

2 Related Work

Despite rapid progress in general question-answering (QA), e.g., (Clark and Gardner, 2018), and formal models for qualitative reasoning (QR), e.g., (Forbus, 1984; Weld and De Kleer, 2013), there has been little work on reasoning with *textual* qualitative knowledge, and no dataset available in this area. Although many datasets include a few qualitative questions, e.g., (Yang et al., 2018; Clark et al., 2018), the only one directly probing

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 5941–5946,
Hong Kong, China, November 3–7, 2019.

Differing Comparatives:		
Q_1	Jan is comparing stars, specifically a small star and the larger Sun. Given the size of each, Jan can tell that the Sun puts out heat that is (A) greater (B) lesser	
K_1	Bigger stars produce more energy, so their surfaces are hotter.	
Discrete Property Values:		
Q_2	What happens to a light car when it has the same power as a heavy car? (A) accelerates faster (B) accelerates slower	
K_2	The smaller its mass is, the greater its acceleration for a given amount of force.	
Numerical Property Values:		
Q_3	Will found water from a source 10m from shore. Eric found water from a source 2m from shore. Whose water likely contains the least nutrients? (A) Will's (B) Eric's	
K_3	Most nutrients are washed into ocean water from land. Therefore, water closer to shore tends to have more nutrients.	
Commonsense Knowledge:		
Q_4	Compared to a box of bricks a box of feathers would be (A) lighter (B) heavier	
K_4	A given volume of a denser substance is heavier than the same volume of a less dense substance.	
Multiple Entities ("Worlds"):		
Q_5	Jimbo liked to work out, while James never did. Which person would have weaker muscles? (A) Jimbo (B) James	
K_5	Muscles that are exercised are bigger and stronger than muscles that are not exercised.	
Complex Stories:		
Q_6	NASA has sent an unmanned probe to survey a distant solar system with four planets. Planet Zorb is farthest from the sun of this solar system, Planet Krakatoa is second farthest, Planet Beanbag is third, and Krypton is the closest. The probe visits the planets in order, first Zorb, then Krakatoa, then Beanbag and finally Krypton. Did the probe have to fly farther in its trip from (A) Zorb to Krakatoa or (B) from Beanbag to Krypton?	
K_6	In general, the farther away from the Sun, the greater the distance from one planets orbit to the next.	

Table 1: Examples of crowdsourced questions Q and corpus knowledge K in QUARTZ, illustrating phenomena.

QR is QuaRel (Tafjord et al., 2019). However, although QuaRel contains 2700 qualitative questions, its underlying qualitative knowledge was specified formally, using a small, fixed ontology of 19 properties. As a result, systems trained on QuaRel are limited to only questions about those properties. Likewise, although the QR community has performed some work on extracting qualitative models from text, e.g., (McFate et al., 2014; McFate and Forbus, 2016), and interpreting questions about identifying qualitative processes, e.g., (Crouse et al., 2018), there is no dataset available for the NLP community to study textual qualitative reasoning. QUARTZ addresses this need.

3 The Task

Examples of QuaRTz questions Q_i are shown in Table 1, along with a sentence K_i expressing the relevant qualitative relationship. The QUARTZ task is to answer the questions given a small (400 sentence) corpus K of general qualitative relationship sentences. Questions are crowdsourced, and the sentences K_i were collected from a larger corpus, described shortly.

Note that the task involves substantially more than matching intensifiers (more/greater/...) between Q_i and K_i. Answers also require some qualitative reasoning, e.g., if the intensifiers are inverted in the question, and entity tracking, to keep track of which entity an intensifier applies to. For example, consider the qualitative sentence and three questions (correct answers bolded):

K_n: People with greater height are stronger.

Q_n: Sue is taller than Joe so Sue is (A) **stronger** (B) weaker

Q'_n: Sue is shorter than Joe so Sue is (A) stronger (B) **weaker**

Q''_n: Sue is shorter than Joe so Joe is (A) **stronger** (B) weaker

Q'_n requires reasoning about intensifers that are flipped with respect to K (shorter $\rightarrow$ weaker), and Q''_n requires entities be tracked correctly (asking about Sue or Joe changes the answer).

4 Dataset Collection

QUARTZ was constructed as follows. First, 400 sentences[2] expressing general qualitative relations were manually extracted by the authors from a large corpus using keyword search ("increase", "faster", etc.). Examples (K_i) are in Table 1.

Second, crowdworkers were shown a seed sentence K_i, and asked to annotate the two properties

[2] In a few cases, a short paragraph rather than sentence was selected, where surrounding context was needed to make sense of the qualitative statement.

being compared using the template below, illustrated using K_2 from Table 1:

"The smaller its mass is, the greater its acceleration for a given amount of force."

What is being compared?			
Less ▼	mass	More ▼	acceleration

They were then asked to author a situated, 2-way multiple-choice (MC) question that tested this relationship, guided by multiple illustrations. Examples of their questions (Q_i) are in Table 1.

Third, a second set of workers was shown an authored question, asked to validate its answer and quality, and asked to annotate how the properties of K_i identified earlier were expressed. To do this, they filled a second template, illustrated for Q_2:

	Appears in question as:	
Relation:	More/Less phrase:	Changed property:
MORE mass:	-	heavy
LESS mass:	-	light
MORE acceleration:	faster	accelerates
LESS acceleration:	slower	accelerates

Finally these workers were asked to generate a new question by "flipping" the original so the answer switched. Flipping means inverting comparatives (e.g., "more" → "less"), values, and other edits as needed to change the answer, e.g.,

K: *More rain causes damper surfaces.*

Q:*More rain causes (A) wetter land (B) drier land*

Q-flipped: **Less** *rain causes (A) wetter land (B) drier land*

Flipped questions are created to counteract the tendency of workers to use the same comparison direction (e.g., "more") in their question as in the seed sentence K_i, potentially answerable by simply matching Q_i and K_i. Flipped questions are more challenging as they demand more qualitative reasoning (Section 7.1).

Questions marked by workers as poor quality were reviewed by the authors and rejected/modified as appropriate. The dataset was then split into train/dev/test partitions such that questions from the same seed K_i were all in the same partition. Statistics are in Table 2.

To determine if the questions are correct and answerable given the general knowledge, a human baseline was computed. Three annotators independently answered a random sample of 100 questions given the supporting sentence K_i for each. The mean score was 95.0%.

# questions Q_i	3864
flip/no flip	1932/1932
positive/negative qualitative influence (QR+/QR-)	2772/1092
train/dev/test	2696/384/784
av Q_i length (sents) min/avg/max	1/1.5/6
av K_i length (sents) min/avg/max	1/1.1/4

Table 2: Statistics of QUARTZ.

5 Models

The QUARTZ task is to answer the questions given the corpus K of qualitative background knowledge. We also examine a "no knowledge" (questions only) task and a "perfect knowledge" task (each question paired with the qualitative sentence K_i it was based on). We report results using two baselines and several strong models built with BERT-large (Devlin et al., 2019) as follows:

1. **Random:** always 50% (2-way MC).
2. **BERT-Sci:** BERT fine-tuned on a large, general set of science questions (Clark et al., 2018).
3. **BERT (IR):** This model performs the full task. First, a sentence K_i is retrieved from K using Q_i as a search query. This is then supplied to BERT as *[CLS] K_i [SEP] question-stem [SEP] answer-option [SEP]* for each option. The [CLS] output token is projected to a single logit and fed through a softmax layer across answer options, using cross entropy loss, the highest being selected. This model is fine-tuned using QUARTZ (only).
4. **BERT (IR upper bound):** Same, but using the ideal (annotated) K_i rather than retrieved K_i.
5. **BERT-PFT (no knowledge):** BERT first fine-tuned ("pre-fine-tuned") on the RACE dataset (Lai et al., 2017; Sun et al., 2019), and then fine-tuned on QUARTZ (questions only, no K, both train and test). Questions are supplied as *[CLS] question-stem [SEP] answer-option [SEP]*.
6. **BERT-PFT (IR):** Same as BERT (IR), except starting with the pre-fine-tuned BERT.

All models were implemented using AllenNLP (Gardner et al., 2018).

6 Results

The results are shown in Table 3, and provide insights into both the data and the models:

1. **The dataset is hard.** Our best model, BERT-PFT (IR), scores only 73.7, over 20 points behind human performance (95.0), suggesting there are significant linguistic and semantic challenges to overcome (Section 7).

Questions→ Model ↓	All Test	No-flip only	Flip only
Baselines:			
Random	50.0	50.0	50.0
BERT-Sci	54.6	76.0	33.2
Models:			
BERT (IR)	64.4	66.3	62.5
(BERT IR upper bound)	(67.7)	(68.1)	(67.3)
BERT-PFT (no knowledge)	68.8	70.4	67.1
BERT-PFT (IR)	**73.7**	77.3	70.2
(BERT-PFT IR upper bound)	(79.8)	(82.1)	(77.6)
Human	95.0		

Table 3: Performance of various models on QUARTZ.

2. **A general science-trained QA system has not learned this style of reasoning.** BERT-Sci only scores 54.6, just a little above random (50.0).
3. **Pre-Fine-Tuning is important.** Fine-tuning only on QUARTZ does signficantly worse (64.4) than pre-fine-tuning on RACE before fine-tuning on QUARTZ (73.7). Pre-fine-tuning appears to teach BERT something about multiple choice questions in general, helping it more effectively fine-tune on QUARTZ.
4. **BERT already "knows" some qualitative knowledge.** Interestingly, BERT-PFT (no knowledge) scores 68.8, significantly above random, suggesting that BERT already "knows" some kind of qualitative knowledge. To rule out annotation artifacts, we we experimented with balancing the distributions of positive and negative influences, and different train/test splits to ensure no topical overlap between train and test, but the scores remained consistent.
5. **BERT can apply general qualitative knowledge to QA, but only partially.** The model for the full task, BERT-PFT (IR) outperforms the no knowledge version (73.7, vs. 68.8), but still over 20 points below human performance. Even given the ideal knowledge (IR upper bound), it is still substantially behind (at 79.8) human performance. This suggests more sophisticated ways of training and/or reasoning with the knowledge are needed.

7 Discussion and Analysis

7.1 Qualitative Reasoning

Can models learn qualitative reasoning from QUARTZ? While QUARTZ questions do not require chaining, 50% involve "flipping" a qualitative relationship (e.g., K: "more X → more Y", Q: "Does less X → less Y?"). Training on just the original crowdworkers' questions, where they chose to flip the knowledge only 10% of the time, resulted in poor (less than random) performance on all the flipped questions. However, training on full QUARTZ, where no-flip and flip were balanced, resulted in similar score for both types of question, suggesting that such a reasoning capability can indeed be learned.

Comparatives:
"warmer" ↔ "increase temperature"
"more difficult" ↔ "slower"
"need more time" ↔ "have lesser amount"
"decreased distance" ↔ "hugged"
"cost increases" ↔ "more costly"
"increase mass" ↔ "add extra"
"more tightly packed" ↔ "add more"
Commonsense Knowledge:
"more land development" ↔ "city grow larger"
"not moving" ↔ "sits on the sidelines"
"caught early" ↔ 'sooner treated"
"lets more light in" ↔ "get a better picture"
"stronger electrostatic force" ↔ "hairs stand up more"
"less air pressure" ↔ "more difficult to breathe"
"more photosynthesis" ↔ "increase sunlight"
Discrete Values:
"stronger acid" ↔ "vinegar" vs. "tap water"
"more energy" ↔ "ripple" vs. "tidal wave"
"closer to Earth" ↔ "ball on Earth" vs. "ball in space"
"mass" ↔ "baseball" vs. "basketball"
"rougher" ↔ "notebook paper" vs. "sandpaper"
"heavier" ↔ "small wagon" vs. "eighteen wheeler"

Table 4: Examples of linguistic and semantic gaps between knowledge K_i (left) and question Q_i (right). A system needs to bridge such gaps for high performance.

7.2 Linguistic Phenomena

From a detailed analysis of 100 randomly sampled questions, the large majority (86%) involved the (overlapping) linguistic and semantic phenomena below, and illustrated in Tables 1 and 4:

1. **Differing comparative expressions** (≈68%) between K_i and Q_i occur in the majority of questions, e.g.,
   "increased altitude" ↔ "higher"
2. **Indirection and Commonsense knowledge** (≈35%) is needed for about 1/3 of the questions to relate K and Q, e.g.,
   "higher temperatures" ↔ "A/C unit broken"
3. **Multiple Worlds** (≈26%): 1/4 of the questions explicitly mention *both* situations being compared, e.g., Q_1 in Table 1. Such questions are known to be difficult because models can easily confuse the two situations (Tafjord et al., 2019).
4. **Numerical property values** (≈11%) require numeric comparison to identify the qualitative relationship, e.g., that "60 years" is older than "30 years".

5. **Discrete property values** (≈7%), often require commonsense to compare, e.g., that a "melon" is larger than an "orange".
6. **Stories** (≈15%): 15% of the questions are 3 or more sentences long, making comprehension more challenging.

This analysis illustrates the richness of linguistic and semantic phenomena in QUARTZ.

7.3 Use of the Annotations

QUARTZ includes a rich set of annotations on all the knowledge sentences and questions, marking the properties being compared, and the linguistic and semantic comparatives employed (Figure 1). This provides a laboratory for exploring semantic parsing approaches, e.g., (Berant et al., 2013; Krishnamurthy et al., 2017), where the underlying qualitative comparisons are extracted and can be reasoned about.

8 Conclusion

Understanding and applying textual qualitative knowledge is an important skill for question-answering, but has received limited attention, in part due the lack of a broad-coverage dataset to study the task. QUARTZ aims to fill this gap by providing the first *open-domain* dataset of qualitative relationship questions, along with the requisite qualitative knowledge and a rich set of annotations. Specifically, QuaRTz removes the requirement, present in all previous qualitative reasoning work, that a fixed set of qualitative relationships be formally pre-specified. Instead, QuaRTz tests the ability of a system to find and apply an arbitrary relationship on the fly to answer a question, including when simple reasoning (arguments, polarities) is required.

As the QUARTZ task involves using a general corpus K of textual qualitative knowledge, a high-performing system would be close to a fully general system where K was much larger (e.g., the Web or a filtered subset), encompassing many more qualitative relationships, and able to answer arbitrary questions of this kind. Scaling further would thus require more sophisticated retrieval over a larger corpus, and (sometimes) chaining across influences, when a direct connection was not described in the corpus. QUARTZ thus provides a dataset towards this end, allowing controlled experiments while still covering a substantial number of textual relations in an open setting. QuaRTz is available at http://data.allenai.org/quartz/.

Acknowledgements

We are grateful to the AllenNLP and Beaker teams at AI2, and for the insightful discussions with other Aristo team members. Computations on beaker.org were supported in part by credits from Google Cloud.

References

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *EMNLP'13*.

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

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

M. Crouse, C. McFate, and K. Forbus. 2018. Learning to build qualitative scenario models from natural language. In *Proc. 31st Int. Workshop on Qualitative Reasoning (QR'18)*.

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

Kenneth D. Forbus. 1984. Qualitative process theory. *Artificial Intelligence*, 24:85–168.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *NLP OSS Workshop at ACL*. (arXiv:1803.07640).

Jayant Krishnamurthy, Pradeep Dasigi, and Matthew Gardner. 2017. Neural semantic parsing with type constraints for semi-structured tables. In *EMNLP'17*.

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

Clifton McFate and Kenneth Forbus. 2016. Scaling up linguistic processing of qualitative processes. In *Proc. 4th Ann. Conf. on Advances in Cognitive Systems*.

Clifton James McFate, Kenneth D Forbus, and Thomas R Hinrichs. 2014. Using narrative function to extract qualitative information from natural language texts. In *AAAI'14*.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2019. Improving machine reading comprehension with general reading strategies. In *NAACL*.

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*.

Daniel S Weld and Johan De Kleer. 2013. *Readings in qualitative reasoning about physical systems*. Morgan Kaufmann.

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*.