QUAREL: A Dataset and Models for Answering Questions about Qualitative Relationships

Oyvind Tafjord, Peter Clark, Matt Gardner, Wen-tau Yih, Ashish Sabharwal
Allen Institute for AI, Seattle, WA
{oyvindt,peterc,mattg,scottyih,ashishs}@allenai.org

Abstract

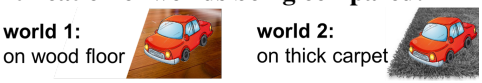
Many natural language questions require recognizing and reasoning with qualitative relationships (e.g., in science, economics, and medicine), but are challenging to answer with corpus-based methods. Qualitative modeling provides tools that support such reasoning, but the semantic parsing task of mapping questions into those models has formidable challenges. We present QUAREL, a dataset of diverse story questions involving qualitative relationships that characterize these challenges, and techniques that begin to address them. The dataset has 2771 questions relating 19 different types of quantities. For example, *"Jenny observes that the robot vacuum cleaner moves slower on the living room carpet than on the bedroom carpet. Which carpet has more friction?"* We contribute (1) a simple and flexible conceptual framework for representing these kinds of questions; (2) the QUAREL dataset, including logical forms, exemplifying the parsing challenges; and (3) two novel models for this task, built as extensions of type-constrained semantic parsing. The first of these models (called QUASP+) significantly outperforms off-the-shelf tools on QUAREL. The second (QUASP+ZERO) demonstrates zero-shot capability, i.e., the ability to handle new qualitative relationships without requiring additional training data, something not possible with previous models. This work thus makes inroads into answering complex, qualitative questions that require reasoning, and scaling to new relationships at low cost. The dataset and models are available at http://data.allenai.org/quarel.

1 Introduction

Many natural language tasks require recognizing and reasoning with qualitative relationships. For example, we may read about temperatures rising (climate science), a drug dose being increased (medicine), or the supply of goods being reduced (economics), and want to reason about the effects. Qualitative story problems, of the kind found in elementary exams (e.g., Figure 1), form a natural example of many of these linguistic and reasoning challenges, and is the target of this work.

Understanding and answering such questions is particularly challenging. Corpus-based methods perform poorly in this setting, as the questions ask about novel scenarios rather than facts that can be looked up. Similarly, word association methods struggle, as a single word change (e.g., "more" to "less") can flip the answer. Rather, the task appears to require *knowledge of the underlying qualitative relations* (e.g., "more friction implies less speed").

Qualitative Story Problem:
Alan noticed that his toy car rolls further on a wood floor than on a thick carpet. This suggests that:
(A) The carpet has more resistance
(B) The floor has more resistance
Solution: (A) The carpet has more resistance
Identification of worlds being compared:



Question Interpretation (Logical Form):
qrel(distance, higher, world1) →
qrel(friction, higher, world2) ;
qrel(friction, higher, world1)?

Figure 1: An example problem from QUAREL and its logical form (LF), from which the answer can be inferred (Section 3). The problem is conceptualized as comparing two *worlds* which the semantic parser needs to identify and track. In the LF, qrel(p, higher|lower, w) denotes that p is higher/lower in world w (compared with the other world). Colors show approximate correspondence between the question and the LF.

Qualitative modeling (Forbus 1984; Weld and De Kleer 2013; Kuipers 1994) provides a means for encoding and reasoning about such relationships. Relationships are expressed in a natural, qualitative way (e.g., if X increases, then so will Y), rather than requiring numeric equations, and inference allows complex questions to be answered. However, the semantic parsing task of mapping real world questions into these models is formidable and presents unique challenges. These challenges must be solved if natural questions involving qualitative relationships are to be reliably answered.

We make three contributions: (1) a simple and flexible conceptual framework for formally representing these kinds of questions, in particular ones that express qualitative comparisons between two scenarios; (2) a challenging new dataset (QUAREL), including logical forms, exemplify-

ing the parsing challenges; and (3) two novel models that extend type-constrained semantic parsing to address these challenges.

Our first model, QUASP+, addresses the problem of tracking different "worlds" in questions, resulting in significantly higher scores than with off-the-shelf tools (Section 7.1). The second model, QUASP+ZERO, demonstrates zero-shot capability, i.e., the ability to handle new qualitative relationships on unseen properties, without requiring additional training data, something not possible with previous models (Section 7.2). Together these contributions make inroads into answering complex, qualitative questions by linking language and reasoning, and offer a new dataset and models to spur further progress by the community.

2 Related Work

There has been rapid progress in question-answering (QA), spanning a wide variety of tasks and phenomena, including factoid QA (Rajpurkar et al. 2016), entailment (Bowman et al. 2015), sentiment (Maas et al. 2011), and ellipsis and coreference (Long, Pasupat, and Liang 2016). Our contribution here is the first dataset specifically targeted at qualitative relationships, an important category of language that has been less explored. While questions requiring reasoning about qualitative relations sometimes appear in other datasets, e.g., (Clark et al. 2018), our dataset specifically focuses on them so their challenges can be studied.

For answering such questions, we treat the problem as mapping language to a structured formalism (semantic parsing) where simple qualitative reasoning can occur. Semantic parsing has a long history (Zelle and Mooney 1996; Zettlemoyer and Collins 2005; Berant et al. 2013; Krishnamurthy, Dasigi, and Gardner 2017), using datasets about geography (Zelle and Mooney 1996), travel booking (Dahl et al. 1994), factoid QA over knowledge bases (Berant et al. 2013), Wikipedia tables (Pasupat and Liang 2015), and many more. Our contributions to this line of research are: a dataset that features phenomena under-represented in prior datasets, namely (1) highly diverse language describing open-domain qualitative problems, and (2) the need to reason over entities that have no explicit formal representation; and methods for adapting existing semantic parsers to address these phenomena.

For the target formalism itself, we draw on the extensive body of work on qualitative reasoning (Forbus 1984; Weld and De Kleer 2013; Kuipers 1994) to create a logical form language that can express the required qualitative knowledge, yet is sufficiently constrained that parsing into it is feasible, described in more detail in Section 3.

There has been some work connecting language with qualitative reasoning, although mainly focused on extracting qualitative models themselves from text rather than question interpretation, e.g., (McFate, Forbus, and Hinrichs 2014; McFate and Forbus 2016). Recent work by Crouse, McFate, and Forbus (2018) also includes interpreting questions that require identifying qualitative processes in text, in constrast to our setting of interpreting NL story questions that involve qualitative comparisons.

Answering story problems has received attention in the domain of arithmetic, where simple algebra story questions (e.g., "Sue had 5 cookies, then gave 2 to Joe...") are mapped to a system of equations, e.g., (Ling et al. 2017; Kushman et al. 2014; Wang, Liu, and Shi 2017; Shi et al. 2015). This task is loosely analogous to ours (we instead map to qualitative relations) except that in arithmetic the entities to relate are often identifiable (namely, the numbers). Our qualitative story questions lack this structure, adding an extra challenge.

The QUAREL dataset shares some structure with the Winograd Schema Challenge (Levesque, Davis, and Morgenstern 2011), being 2-way multiple choice questions invoking both commonsense and coreference. However, they test different aspects of commonsense: Winograd uses coreference resolution to test commonsense understanding of scenarios, while QUAREL tests reasoning about qualitative relationships requiring tracking of coreferent "worlds."

Finally, crowdsourcing datasets has become a driving force in AI, producing significant progress, e.g., (Rajpurkar et al. 2016; Joshi et al. 2017; Wang, Berant, and Liang 2015). However, for semantic parsing tasks, one obstacle has been the difficulty in crowdsourcing target logical forms for questions. Here, we show how those logical forms can be obtained indirectly from workers without training the workers in the formalism, loosely similar to (Yih et al. 2016).

3 Knowledge Representation

We first describe our framework for representing questions and the knowledge to answer them. Our dataset, described later, includes logical forms expressed in this language.

3.1 Qualitative Background Knowledge

We use a simple representation of qualitative relationships, leveraging prior work in qualitative reasoning (Forbus 1984). Let $P = \{p_i\}$ be the set of properties relevant to the question set's domain (e.g., smoothness, friction, speed). Let $V_i = \{v_{ij}\}$ be a set of qualitative values for property p_i (e.g., fast, slow). For the *background knowledge* about the domain itself (a qualitative model), following Forbus (1984), we use the following predicates:

q+(*property1*, *property2*)
q-(*property1*, *property2*)

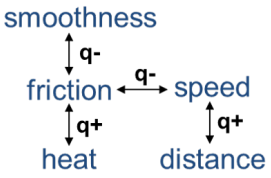
q+ denotes that *property1* and *property2* are qualitatively proportional, e.g., if *property1* goes up, *property2* will too, while q- denotes inverse proportionality, e.g.,

 # If friction goes up, speed goes down.
q-(friction, speed).

We also introduce the predicate:

higher-than(val_{ij}, val_{ik}, *property*$_i$)

where $val_{ij} \in V_i$, allowing an ordering of property values to be specified, e.g., higher-than(fast, slow, speed). For our purposes here, we simplify to use just two property values, low and high, for all properties. (The parser learns mappings from words to these values, described later).



q-(smoothness,friction).
q+(friction,heat).
q-(friction,speed).
q+(distance,speed).

Figure 2: A simple qualitative theory about friction, shown graphically (left) and formally (right). For example, q-(smoothness,friction) indicates that if smoothness increases, friction decreases.

Given these primitives, compact theories can be authored for a particular domain by choosing relevant properties P, and specifying qualitative relationships (q+,q-) and ordinal values (higher-than) for them. For example, a simple theory about friction is sketched graphically in Figure 2. Our observation is that these theories are relatively small, simple, and easy to author. Rather, the primary challenge is in mapping the complex and varied language of questions into a form that interfaces with this representation.

This language can be extended to include additional primitives from qualitative modeling, e.g., i+(x,y) ("the *rate of change* of x is qualitatively proportional to y"). That is, the techniques we present are not specific to our particular qualitative modeling subset. The only requirement is that, given a set of absolute values or qualitative relationships from a question, the theory can compute an answer.

3.2 Representing Questions

Predicates. A key feature of our representation is the conceptualization of questions as describing events happening in two *worlds*, world1 and world2, that are being compared. That comparison may be between two different entities, or the same entity at different time points. E.g., in Figure 1 the two worlds being compared are the car on wood, and the car on carpet. The tags world1 and world2 denote these different situations, and semantic parsing (Section 5) requires learning to correctly associate these tags with parts of the question describing those situations. This abstracts away irrelevant details of the worlds, while still keeping track of which world is which.

We define the following two predicates to express qualitative information in questions:

qrel(*property, direction, world*)
qval(*property, value, world*)

where *property* (p_i) $\in$ P, *value* $\in V_i$, *direction* $\in$ {higher, lower}, and *world* $\in$ {world1, world2}. **qrel()** denotes the relative assertion that *property* is *higher/lower* in *world* compared with the other world, which is left implicit,[1] e.g., from Figure 1:

 # The car rolls further on wood.
qrel(distance, higher, world1)

where world1 is a tag for the "car on wood" situation (hence world2 becomes a tag for the opposite "car on carpet" situation). **qval()** denotes that *property* has an absolute *value* in *world*, e.g.,

 # The car's speed is slow on carpet.
qval(speed, low, world2)

3.3 Logical Forms for Questions

Despite the wide variation in language, the space of logical forms (LFs) for the questions that we consider is relatively compact. In each question, the question body establishes a scenario and each answer option then probes an implication. We thus express a question's LF as a tuple:

(setup, answer-A, answer-B)

where *setup* is the predicate(s) describing the scenario, and *answer-** are the predicate(s) being queried for. If *answer-A* follows from *setup*, as inferred by the reasoner, then the answer is (A); similarly for (B). For readability we will write this as

setup $\rightarrow$ *answer-A* ; *answer-B*

We consider two styles of LF, covering a large range of questions. The first is:

(1) qrel(p, d, w) $\rightarrow$
qrel(p', d', w') ; qrel(p', d'', w'')

which deals with relative values of properties between worlds, and applies when the question setup includes a comparative. An example of this is in Figure 1. The second is:

(2) qval(p, v, w), qval(p, v', w''') $\rightarrow$
qrel(p', d', w') ; qrel(p', d'', w'')

which deals with absolute values of properties, and applies when the setup uses absolute terms instead of comparatives. An example is the first question in Figure 3, shown simplified below, whose LF looks as follows (colors showing approximate correspondences):

 # Does a bar stool slide faster along the bar surface with decorative raised bumps or the smooth wooden floor? (A) bar (B) floor
qval(smoothness, low, world1),
qval(smoothness, high, world2) $\rightarrow$
qrel(speed, higher, world1) ;
qrel(speed, higher, world2)

3.4 Inference

A small set of rules for qualitative reasoning connects these predicates together. For example, (in logic) **if** the value of P is higher in world1 than the value of P in world2 **and** q+(P,Q) **then** the value of Q will be higher in world1 than the value of Q in world2. Given a question's logical form, a qualitative model, and these rules, a Prolog-style inference engine determines which answer option follows from the premise.[2]

[1] We consider just two worlds being compared here, but the formalism generalizes to N-way comparisons by adding a fourth argument: qrel(*prop, dir, world, other-world*).

[2] E.g., in Figure 1, the qualitative model includes q-(friction, distance), and the general qualitative reasoning rules include oppo-

(1) Heather wants to see if a bar stool will slide faster along the bar surface which has decorative raised bumps on it or on the smooth wooden floor. On which surface will the chair slide faster? (A) bar (B) floor
LF: *qval(smoothness, low, world1), qval(smoothness, high, world2) → qrel(speed, higher, world1) ; qrel(speed, higher, world2)*

(2) Andy was running across the tile floor and sliding across it. There was less friction here, but he thought he could do the same outside on the cement. When Andy tries to slide across the cement, his socks will make _____ than when he slides across the tile floor. (A) more heat (B) less heat
LF: *qrel(friction, lower, world1) → qrel(heat, higher, world2 ; qrel(heat, lower, world2)*

(3) Mary noticed that erasing her mistakes on her drawing paper seemed to take more effort than the marker paper. This caused more heat to develop on (A) the drawing paper or (B) the marker paper
LF: *qrel(friction, higher, world1) → qrel(heat, higher, world1) ; qrel(heat, higher, world2)*

(4) Henry is playing with his younger brother. Henry is bigger and stronger and he can throw the ball (A) farther (B) not as far.
LF: *qrel(strength, higher, world1) → qrel(distance, higher, world1) ; qrel(distance, lower, world1)*

(5) It's turkey hunting season and Jim is on his front porch. He hears a gun shot off the the west. Then he hears another one off to the north. The one to the north was easier to hear than the one to the west. Which hunter is closer to Jim's house? (A) the one to the west (B) the one to the north
LF: *qrel(loudness, higher, world1) → qrel(distance, lower, world2) ; qrel(distance, lower, world1)*

Figure 3: Examples of questions and logical forms in the QUAREL dataset (the first 3 are also in the friction subset, QUAREL^F)

4 The QUAREL Dataset

QUAREL is a crowdsourced dataset of 2771 multiple-choice story questions, including their logical forms. The size of the dataset is similar to several other datasets with annotated logical forms used for semantic parsing (Zelle and Mooney 1996; Hemphill, Godfrey, and Doddington 1990; Yih et al. 2016). As the space of LFs is constrained, the dataset is sufficient for a rich exploration of this space.

We crowdsourced multiple-choice questions in two parts, encouraging workers to be imaginative and varied in their use of language. First, workers were given a seed qualitative relation q+/-(p_1, p_2) in the domain, expressed in English (e.g., "If a surface has more friction, then an object will travel slower"), and asked to enter two objects, people, or situations to compare. They then created a question, guided by a large number of examples, and were encouraged to be imaginative and use their own words. The results are a remarkable variety of situations and phrasings (Figure 3).

Second, the LFs were elicited using a novel technique of reverse-engineering them from a set of follow-up questions, without exposing workers to the underlying formalism. This is possible because of the constrained space of LFs. Referring to LF templates (1) and (2) earlier (Section 3.3), these questions are as follows:

1. What is the correct answer (A or B)?
2. Which property are the answer options asking about? ($p' \in \{p_1, p_2\}$)
3. In the correct answer, is this property higher or lower than in the incorrect answer? (d')
4. Do the answer options:
   - ask the *same* question about *different* objects/situations? ($d' = d'', w' \neq w''$)
   - ask *opposite* questions about *the same* object/situation? ($d' \neq d'', w' = w''$)
5. Which direction of comparison is used in the body of the question?
   - higher/lower? (d, LF template is (1))
   - *OR* were two values given? If so, enter the values, standardized as high/low in the LF (v,v', LF template is (2))

From this information, we can deduce the target LF (p is the complement of $p' \in \{p_1, p_2\}$, $w''' \neq w$, we arbitrarily set w=world1, hence all other variables can be inferred). Three independent workers answer these follow-up questions to ensure reliable results.

We also had a human answer the questions in the dev partition (in principle, they should all be answerable). The human scored 96.4%, the few failures caused by occasional annotation errors or ambiguities in the question set itself, suggesting high fidelity of the content.

About half of the dataset are questions about friction, relating five different properties (friction, heat, distance, speed, smoothness). These questions form a meaningful, connected subset of the dataset which we denote QUAREL^F. The remaining questions involve a wide variety of 14 additional properties and their relations, such as "exercise intensity vs. sweat" or "distance vs. brightness".[3]

Figure 3 shows typical examples of questions in QUAREL, and Table 1 provides summary statistics. In particular, the vocabulary is highly varied (5226 unique words), given the dataset size. Figure 4 shows some examples of the varied phrases used to describe smoothness.

5 Baseline Systems

We use four systems to evaluate the difficulty of this dataset. (We subsequently present two new models, extending the baseline neural semantic parser, in Sections 7.1 and 7.2).

site(world1, world2) and qrel(P, D, W) ∧ q-(P, P') ∧ opposite(W, W') → qrel(P', D, W'), so the answer can be inferred.

[3] See supplementary material at http://data.allenai.org/quarel for a complete list.

Property	Size
# questions	2771
# questions (QUAREL^F subset)	1395
# type (1), (2) (see Section 3.3)	2048, 723
# questions train/dev/test	1941/278/552
Min/avg/max qn length (words)	11/37/112
Min/avg/max qn length (sent.)	1/2.4/8
Vocab (# uniq words)	5226

Table 1: Summary statistics for the QUAREL dataset.

"smoother" (864), "rougher" (568), "more smooth" (270), "more rough" (126), "bumpier" (55), "less bumpy" (27), "more rugged" (24), "easier" (16), "not as rough" (16), "flatter" (10), "slicker" (10), "isn't as rough" (8), "stickier" (8), "rugged" (7), "more even" (6), "more easily" (5), "spikier" (4), "more uniform" (4), "softer" (4), "more level" (4), "bumpier ride" (4), "glide more easily" (4), "more jagged" (4), "lumpier" (3), "more fluidly" (3), "rolls easily" (2), "struggle to move" (2), "ideal surface" (2), "more freely" (2), "harder" (2), "has more ridges" (2), "more uniform surface" (2), "more scraping" (2), "more ridges" (2), "sleek" (2), "easier to pull" (1), "barely a touch" (1), "flatter surface" (1), "much better" (1), "full of bumps" (1), "rocky" (1), "calmer" (1), "less slick" (1), "bumpiness" (1), "less obstacles" (1), "more silky glide" (1), "sanded" (1), "slick" (1), "lots of bumps" (1), "not nearly as smooth" (1), "easier to wipe" (1), "snagging" (1), "easier on my feet" (1), "rolls more easily" (1)

Figure 4: Examples of the varied way that smoother/rougher surfaces are described in QUAREL questions.

The first two are an information retrieval system and a word-association method, following the designs of Clark et al. (2016). These are naive baselines that do not parse the question, but nevertheless may find some signal in a large corpus of text that helps guess the correct answer. The third is a CCG-style rule-based semantic parser written specifically for friction questions (the QUAREL^F subset), but prior to data being collected. The last is a state-of-the-art neural semantic parser. We briefly describe each in turn.

Information Retrieval (IR) System To answer multiple-choice questions, this system searches a large (280GB) text corpus to see if the question q along with an answer option is loosely stated in the corpus, and returns the confidence that such a statement was found. To do this, for each answer option a_i, it sends $q + a_i$ as a query to a search engine and returns the search engine's score for the top retrieved sentence s where s also has at least one non-stopword overlap with q, and at least one with a_i. The option with the highest score is selected.

Pointwise Mutual Information (PMI) Word co-occurrences may also provide some signal for answering these questions, e.g., the high co-occurrence of "faster" and "ice" in a corpus may help answer a question ending with "...faster? (A) ice (B) gravel". To formalize this, given a question q and an answer option a_i, we use PMI (Church and Hanks 1989) to measure the strength of the associations between parts of q and parts of a_i. Given a large corpus C,

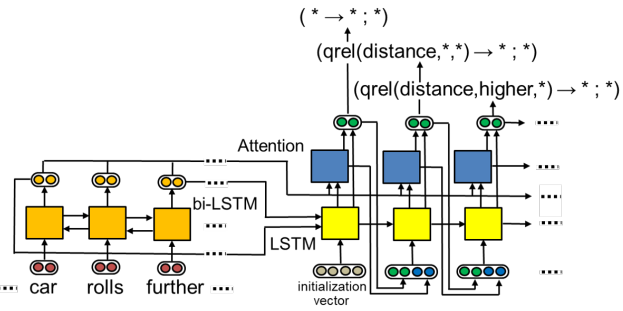


Figure 5: The QUASP parser decodes to a sequence of LF-building decisions, incrementally constructing the LF by selecting production rules from the LF grammar. As illustrated, first it decides if the LF should be of type 1 or 2 (here, type 1 is chosen), then it selects the the property for the question body (here, distance), then it selects the direction of change (here, higher), and so on.

the PMI for two n-grams x and y is defined as

$$\text{PMI}(x, y) = \log \frac{p(x, y)}{p(x)p(y)}$$

The system selects the answer with the largest average PMI, calculated over all pairs of question n-grams and answer option n-grams.

Rule-based Semantic Parser The rule-based semantic parser uses a simple CCG-like grammar (Steedman and Baldridge 2011) specifically written for the friction scenario task (QUAREL^F) over several days, but prior to the dataset being constructed. It represents a good-faith attempt to solve this subset of questions with traditional methods. First, the question is preprocessed to tag likely references to the worlds being compared, using hand-written rules that look for surface names ("road", "ice"), appropriate adjectives ("rough", "green"), and by position ("over <X>"). The first candidate word/phrase is tagged world1 (with type WORLD), the second world2, and if those phrases occur later in the question, they are tagged with the corresponding world. The system then parses the question using 142 task-specific, CCG-like rules, such as:

"is greater than" ⊢ (S\PROPERTY)\WORLD:
λp.λw.qrel(p, HIGHER, w)
"velocity" ⊢ PROPERTY:speed

where \WORLD means "look left for something of category WORLD". Thus a tagged phrase like

"the velocity on ice[world2] is greater than"

produces qrel(speed, higher, world2). The parser skips over most words in the story, only paying attention to words that are tagged or referenced in the grammar.

Type-constrained Neural Semantic Parser (QUASP) Our final system is AllenNLP's implementation of a neural semantic parser (Gardner et al. 2018). This parser uses a

type-constrained, encoder-decoder architecture, representative of the current state-of-the-art on many datasets (Krishnamurthy, Dasigi, and Gardner 2017; Yin and Neubig 2017; Goldman et al. 2018). The model architecture is similar to standard seq2seq models, with an LSTM that encodes the question and an LSTM with attention over the encoded question that decodes a logical form. However, unlike standard seq2seq models that output logical form tokens directly, this parser outputs production rules from a CFG-like grammar over the space of all logical forms. These production rules sequentially build up an abstract syntax tree, which determines the logical form. In this way, the parser is constrained to only produce valid LFs, and does not have to spend modeling capacity learning the syntax of the language.

For our domain, we created a simple grammar capturing the logical form language described in Section 3.3. The parser uses this grammar to find the set of valid choices at each step of decoding. The model architecture, with example inputs and outputs, is illustrated in Figure 5. We refer to this instantiation of the parser as QUASP. As QUAREL has annotated logical forms, this model is trained to maximize the likelihood of the logical form associated with each question. At test time, beam search is used to find the highest scoring parse.

As input to the model we feed the full question plus answer options as a single sequence of tokens, encoding each token using a concatenation of Glove (Pennington, Socher, and Manning 2014) and ELMo (Peters et al. 2018) vectors.

As a separate baseline, we also train a similar two-layer bi-directional LSTM encoder (BILSTM in the results) to directly predict answer A vs. B, without an intermediate logical form.[4]

6 Baseline Experiments

We ran the above systems on the QUAREL dataset. QUASP was trained on the training set, using the model with highest parse accuracy on the dev set (similarly BILSTM used highest answer accuracy on the dev set) . The results are shown in Table 2. The 95% confidence interval is +/- 4% on the full test set. The human score is the sanity check on the dev set (Section 4).

As Table 2 shows, the QUASP model performs better than other baseline approaches which are only slightly above random. QUASP scores 56.1% (61.7% on the friction subset), indicating the challenges of this dataset.

For the rule-based system, we observe that it is unable to parse the majority (66%) of questions (hence scoring 0.5 for those questions, reflecting a random guess), due to the varied and unexpected vocabulary present in the dataset. For example, Figure 4 shows some of the ways that the notion of "smoother/rougher" is expressed in questions, many of which are not covered by the hand-written CCG grammar. This reflects the typical brittleness of hand-built systems.

For QUASP, we also analyzed the parse accuracies, shown in Table 3, the score reflecting the percentage of times

[4]For more implementation details, see supplementary material at http://data.allenai.org/quarel.

Dataset →	**QUAREL**		**QUAREL^F**	
Model ↓	**Dev**	**Test**	**Dev**	**Test**
Random	50.0	50.0	50.0	50.0
Human	96.4	-	95.0	-
IR	50.7	48.6	50.7	48.9
PMI	49.3	50.5	50.7	52.5
Rule-Based	-	-	55.0	57.7
BILSTM	55.8	53.1	59.3	54.3
QUASP	62.1	56.1	69.2	61.7
QUASP+	**68.9**	**68.7**	**79.6**	**74.5**

Table 2: Scores (answer accuracies) of the different models on the full QUAREL dataset and QUAREL^F subset about friction. The baseline models only marginally outperform a random baseline. In QUASP+, however, identifying and delexicalizing the worlds significantly improves the performance (see Section 7.1).

Dataset →	**QUAREL**		**QUAREL^F**	
Model ↓	**Dev**	**Test**	**Dev**	**Test**
QUASP	37.4	32.2	47.9	43.2
QUASP+	**46.8**	**43.8**	**64.3**	**59.0**

Table 3: Parse accuracies for the semantic parsers.

it produced exactly the right logical form. The random baseline for parse accuracy is near zero given the large space of logical forms, while the model parse accuracies are relatively high, much better than a random baseline.

Further analysis of the predicted LFs indicates that the neural model does well at predicting the properties (∼25% of errors on dev set), but struggles to predict the worlds in the LFs reliably (∼70% of errors on dev set). This helps explain why non-trivial parse accuracy does not necessarily translate into correspondingly higher answer accuracy: If only the world assignment is wrong, the answer will flip and give a score of zero, rather than the average 0.5.

7 New Models

We now present two new models, both extensions of the neural baseline QUASP. The first, QUASP+, addresses the leading cause of failure just described, namely the problem of identifying the two worlds being compared, and significantly outperforms all the baseline systems. The second, QUASP+ZERO, addresses the scaling problem, namely the costly requirement of needing many training examples each time a new qualitative property is introduced. It does this by instead using only a small amount of lexical information about the new property, thus achieving "zero shot" performance, i.e., handling properties unseen in the training examples (Palatucci et al. 2009), a capability not present in the baseline systems. We present the models and results for each.

7.1 QUASP+: A Model Incorporating World Tracking

We define the world tracking problem as identifying and tracking references to different "worlds" being compared in

text, i.e., correctly mapping phrases to world identifiers, a critical aspect of the semantic parsing task. There are three reasons why this is challenging. First, unlike properties, the worlds being compared in questions are *distinct in almost every question*, and thus there is no obvious, learnable mapping from phrases to worlds. For example, while a property (like speed) has learnable ways to refer to it ("faster", "moves rapidly", "speeds", "barely moves"), worlds are different in each question (e.g., "on a road", "countertop", "while cutting grass") and thus learning to identify them is hard. Second, different phrases may be used to refer to the same world in the same question (see Figure 6), further complicating the task. Finally, even if the model could learn to identify worlds in other ways, e.g., by syntactic position in the question, there is the problem of selecting world1 or world2 consistently throughout the parse, so that the equivalent phrasings are assigned the same world.

This problem of mapping phrases to world identifiers is similar to the task of entity linking (Ling, Singh, and Weld 2015). In prior semantic parsing work, entity linking is relatively straightforward: simple string-matching heuristics are often sufficient (Jia and Liang 2016; Dong and Lapata 2016), or an external entity linking system can be used (Yih et al. 2015; Xu et al. 2016). In QUAREL, however, because the phrases denoting world1 and world2 are different in almost every question, and the word "world" is never used, such methods cannot be applied.

To address this, we have developed QUASP+, a new model that extends QUASP by adding an extra initial step to identify and delexicalize world references in the question. In this delexicalization process, potentially new linguistic descriptions of worlds are replaced by *canonical tokens*, creating the opportunity for the model to generalize across questions. For example, the world mentions in the question:

> "A ball rolls further on wood than carpet because the (A) carpet is smoother (B) wood is smoother"

are delexicalized to:

> "A ball rolls further on WORLD1 than WORLD2 because the (A) WORLD2 is smoother (B) WORLD1 is smoother"

This approach is analogous to Herzig and Berant (2018), who delexicalized words to POS tags to avoid memorization. Similar delexicalized features have also been employed in Open Information Extraction (Etzioni et al. 2008), so the Open IE system could learn a *general* model of how relations are expressed. In our case, however, delexicalizing to WORLD1 and WORLD2 is itself a significant challenge, because identifying phrases referring to worlds is substantially more complex than (say) identifying parts of speech.

To perform this delexicalization step, we use the world annotations included as part of the training dataset (Section 4) to train a separate tagger to identify "world mentions" (text spans) in the question using BIO tags[5] (BiLSTM encoder followed by a CRF). The spans are then sorted into WORLD1 and WORLD2 using the following algorithm:

[5]e.g., the world mention "calm water" in the question "...in calm water, but..." would be tagged "...in/O calm/B water/I, but/O..."

"road" & "paved roadway"
"wooden bar" & "wood counter"
"her counter is stone" & "stone counter"
"grass" & "(mowing his) yard"
"shag carpeting" & "carpet"
"tiled floor" & "tile"
"wet tennis court" & "wet court"
"wastebasket" & "waste basket"
"ice on the pond" & "ice pond"
"wood beam" & "wooden beam"
"outside" & "grass"
"street" & "asphalt"
"carpet" & "carpeted floor"
"hardwood" & "wood"
"beach" & "sand"
"mulch" & "mulched area"

Figure 6: Examples of different linguistic expressions of the same world in a question.

1. If one span is a substring of another, they are are grouped together. Remaining spans are singleton groups.
2. The two groups containing the longest spans are labeled as the two worlds being compared.
3. Any additional spans are assigned to one of these two groups based on closest edit distance (or ignored if zero overlap).
4. The group appearing first in the question is labeled WORLD1, the other WORLD2.

The result is a question in which world mentions are canonicalized. The semantic parser QUASP is then trained using these questions.[6] We call the combined system (delexicalization plus semantic parser) QUASP+.

The results for QUASP+ are included in Table 2. Most importantly, QUASP+ significantly outperforms the baselines by over 12% absolute. Similarly, the parse accuracies are significantly improved from 32.2% to 43.8% (Table 3). This suggests that this delexicalization technique is an effective way of making progress on this dataset, and more generally on problems where multiple situations are being compared, a common characteristic of qualitative problems.

7.2 QUASP+ZERO: A Model for the Zero-Shot Task

While our delexicalization procedure demonstrates a way of addressing the world tracking problem, the approach still relies on annotated data; if we were to add new qualitative relations, new training data would be needed, which is a significant scalability obstacle. To address this, we define the zero-shot problem as being able to answer questions involving a new predicate p given training data only about other predicates P different from p. For example, if we add a new property (e.g., heat) to the qualitative model (e.g., adding q+(friction, heat); "more friction implies more heat"), we want to answer questions involving heat without creating

[6]During training, using alignment with the annotations, we ensure the worlds in the LF are numbered consistently with these tags.

new annotated training questions, and instead only use minimal extra information about the new property. A parser that achieved good zero-shot performance, i.e., worked well for new properties unseen at training time, would be a substantial advance, allowing a new qualitative model to link to questions with minimal effort.

QUAREL provides an environment in which methods for this zero-shot theory extension can be devised and evaluated. To do this, we consider the following experimental setting: All questions mentioning a particular property are removed, the parser is trained on the remainder, and then tested on those withheld questions, i.e., questions mentioning a property unseen in the training data.

We present and evaluate a model that we have developed for this, called QUASP+ZERO, that modifies the QUASP+ parser as follows: During decoding, at points where the parser is selecting which property to include in the LF (e.g., Figure 5), it does not just consider the question tokens, but also the *relationship* between those tokens and the properties P used in the qualitative model. For example, a question token such as "longer" can act as a cue for (the property) length, even if unseen in the training data, because "longer" and a lexical form of length (e.g.,"length") are similar. This approach follows the entity-linking approach used by Krishnamurthy, Dasigi, and Gardner (2017), where the similarity between question tokens and (words associated with) entities - called the entity linking score - help decide which entities to include in the LF during parsing. Here, we modify their entity linking score $s(p, i)$, linking question tokens q_i and property "entities" p, to be:

$$s(p, i) = \max_{w \in W(p)} v_w^T K v_{q_i}$$

where K is a diagonal matrix connecting the embedding of the question token q_i and words $W(p)$ associated with the property p. For $W(p)$, we provide a small list of words for each property (such as "speed", "velocity", and "fast" for the speed property), a small-cost requirement.

The results with QUASP+ZERO are in Table 4, shown in detail on the QUAREL^F subset and (due to space constraints) summarized for the full QUAREL. We can measure overall performance of QUASP+ZERO by averaging each of the zero-shot test sets (weighted by the number of questions in each set), resulting in an overall parse accuracy of 38.9% and answer accuracy 61.0% on QUAREL^F, and 25.7% (parse) and 59.5% (answer) on QUAREL, both significantly better than random. These initial results are encouraging, suggesting that it may be possible to parse into modified qualitative models that include new relations, with minimal annotation effort, significantly opening up qualitative reasoning methods for QA.

Held out property	Parse (LF) Accuracy Seen	Parse (LF) Accuracy Unseen	Answer Accuracy Seen	Answer Accuracy Unseen
distance	46.7	33.3	67.0	59.8
friction	54.4	29.7	78.6	59.4
heat	33.5	52.6	58.8	64.7
smoothness	47.9	53.2	67.7	62.2
speed	45.0	33.1	64.9	60.2
None	51.8	NA	66.7	NA
Weighted avg.	44.5	**38.9**	66.2	**61.0**

Table 4: Baseline scores (bold) using QUASP+ZERO for the zero-shot task of answering questions involving properties unseen in the training data, using the QUAREL^F subset of QUAREL. For the entire QUAREL dataset, the weighted average scores for questions with unseen properties are **25.7%** (parse) and **59.5%** (answer).

8 Summary and Conclusion

Our goal is to answer questions that involve qualitative relationships, an important genre of task that involves both language and knowledge, but also one that presents significant challenges for semantic parsing. To this end we have developed a simple and flexible formalism for representing these questions; constructed QUAREL, the first dataset of qualitative story questions that exemplifies these challenges; and presented two new models that adapt existing parsing techniques to this task. The first model, QUASP+, illustrates how delexicalization can help with world tracking (identifying different "worlds" in questions), resulting in state-of-the-art performance on QUAREL. The second model, QUASP+ZERO, illustrates how zero-shot learning can be achieved (i.e., adding new qualitative relationships without requiring new training examples) by using an entity-linking approach applied to properties - a capability not present in previous models.

There are several directions in which this work can be expanded. First, quantitative property values (e.g., "10 mph") are currently not handled well, as their mapping to "low" or "high" is context-dependent. Second, some questions do not fit our two question templates (Section 3.3), e.g., where two property values are a single answer option (e.g., "....(A) one floor is smooth and the other floor is rough"). Finally, some questions include an additional level of indirection, requiring an inference step to map to qualitative relations. For example, "Which surface would be best for a race? (A) gravel (B) blacktop" requires the additional commonsense inference that "best for a race" implies "higher speed".

Given the ubiquity of qualitative comparisons in natural text, recognizing and reasoning with qualitative relationships is likely to remain an important task for AI. This work makes inroads into this task, and contributes a dataset and models to encourage progress by others. The dataset and models are publicly available at http://data.allenai.org/quarel.

References

Berant, J.; Chou, A.; Frostig, R.; and Liang, P. 2013. Semantic parsing on Freebase from question-answer pairs. In *EMNLP'13*.

Bowman, S.; Angeli, G.; Potts, C.; and Manning, C. 2015. A large annotated corpus for learning natural language inference. In *EMNLP'15*.

Church, K. W., and Hanks, P. 1989. Word association norms, mutual information and lexicography. In *ACL'89*.

Clark, P.; Etzioni, O.; Khot, T.; Sabharwal, A.; Tafjord, O.; Turney, P. D.; and Khashabi, D. 2016. Combining retrieval, statistics, and inference to answer elementary science questions. In *AAAI'16*.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

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

Dahl, D. A.; Bates, M.; Brown, M.; Fisher, W.; Hunicke-Smith, K.; Pallett, D.; Pao, C.; Rudnicky, A.; and Shriberg, E. 1994. Expanding the scope of the ATIS task: The ATIS-3 corpus. In *HLT'94*.

Dong, L., and Lapata, M. 2016. Language to logical form with neural attention. In *ACL'16*.

Etzioni, O.; Banko, M.; Soderland, S.; and Weld, D. S. 2008. Open information extraction from the web. *Commun. ACM* 51(12):68–74.

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

Gardner, M.; Grus, J.; Neumann, M.; Tafjord, O.; Dasigi, P.; Liu, N.; Peters, M.; Schmitz, M.; and Zettlemoyer, L. 2018. AllenNLP: A deep semantic natural language processing platform. In *NLP OSS Workshop at ACL*. (arXiv:1803.07640).

Goldman, O.; Latcinnik, V.; Naveh, U.; Globerson, A.; and Berant, J. 2018. Weakly-supervised semantic parsing with abstract examples. In *ACL'18*.

Hemphill, C. T.; Godfrey, J. J.; and Doddington, G. R. 1990. The ATIS spoken language systems pilot corpus. In *HLT'90*.

Herzig, J., and Berant, J. 2018. Decoupling structure and lexicon for zero-shot semantic parsing. *CoRR* abs/1804.07918.

Jia, R., and Liang, P. 2016. Data recombination for neural semantic parsing. In *ACL'16*.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL'17*.

Krishnamurthy, J.; Dasigi, P.; and Gardner, M. 2017. Neural semantic parsing with type constraints for semi-structured tables. In *EMNLP'17*.

Kuipers, B. 1994. *Qualitative reasoning: modeling and simulation with incomplete knowledge*. MIT press.

Kushman, N.; Zettlemoyer, L. S.; Barzilay, R.; and Artzi, Y. 2014. Learning to automatically solve algebra word problems. In *ACL'14*.

Levesque, H. J.; Davis, E.; and Morgenstern, L. 2011. The winograd schema challenge. In *AAAI'11*.

Ling, W.; Yogatama, D.; Dyer, C.; and Blunsom, P. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *ACL'11*.

Ling, X.; Singh, S.; and Weld, D. S. 2015. Design challenges for entity linking. *TACL* 3:315–328.

Long, R.; Pasupat, P.; and Liang, P. 2016. Simpler context-dependent logical forms via model projections. In *ACL'16*.

Maas, A. L.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning word vectors for sentiment analysis. In *ACL'11*.

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

McFate, C. J.; Forbus, K. D.; and Hinrichs, T. R. 2014. Using narrative function to extract qualitative information from natural language texts. In *AAAI'14*.

Palatucci, M.; Pomerleau, D.; Hinton, G. E.; and Mitchell, T. M. 2009. Zero-shot learning with semantic output codes. In *NIPS'09*.

Pasupat, P., and Liang, P. 2015. Compositional semantic parsing on semi-structured tables. In *ACL'15*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global vectors for word representation. In *EMNLP'14*.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *NAACL-HLT'18*.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP'16*.

Shi, S.; Wang, Y.; Lin, C.-Y.; Liu, X.; and Rui, Y. 2015. Automatically solving number word problems by semantic parsing and reasoning. In *EMNLP'15*.

Steedman, M., and Baldridge, J. 2011. Combinatory categorial grammar. *Non-Transformational Syntax: Formal and Explicit Models of Grammar* 181–224.

Wang, Y.; Berant, J.; and Liang, P. 2015. Building a semantic parser overnight. In *ACL'15*.

Wang, Y.; Liu, X.; and Shi, S. 2017. Deep neural solver for math word problems. In *EMNLP'17*.

Weld, D. S., and De Kleer, J. 2013. *Readings in qualitative reasoning about physical systems*. Morgan Kaufmann.

Xu, K.; Reddy, S.; Feng, Y.; Huang, S.; and Zhao, D. 2016. Question answering on freebase via relation extraction and textual evidence. In *ACL'16*.

Yih, W.-t.; Chang, M.-W.; He, X.; and Gao, J. 2015. Semantic parsing via staged query graph generation: Question answering with knowledge base. In *ACL'15*.

Yih, W.-t.; Richardson, M.; Meek, C.; Chang, M.-W.; and Suh, J. 2016. The value of semantic parse labeling for knowledge base question answering. In *ACL'16*.

Yin, P., and Neubig, G. 2017. A syntactic neural model for general-purpose code generation. In *ACL'17*.

Zelle, J. M., and Mooney, R. J. 1996. Learning to parse database queries using inductive logic programming. In *AAAI'96*.

Zettlemoyer, L. S., and Collins, M. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *UAI'05*.