Towards High-Precision Understanding of Comparative Analysis Problems Expressed in Natural Language

Maxwell Crouse and Kenneth Forbus[1]

Abstract. The naturalness of qualitative representations suggests that they have an important role to play in natural language semantics. Prior work has focused on extracting qualitative models from text, but very little work has been done on using qualitative reasoning in answering questions posed in natural language. The recent QuaRel dataset developed by AI2 provides a novel opportunity to explore understanding and answering comparative analysis questions expressed in natural language. While some machine learning models have been built at varying levels of performance, none have achieved human-level performance on this task, and we expect that their performance is actually brittle, susceptible to adversarial attacks. This paper proposes an alternate approach, using relational representations, including qualitative relations, to achieve high-precision understanding of the language used in such problems. We describe our approach and progress on quantity identification in this task.

1 Introduction

Comparative analysis [1] uses qualitative representations to ascertain the causal consequences of differences. This includes the effects of a hypothetical change to a system, e.g. using a stiffer spring in a mechanical design, or the differences between two physical systems, e.g. the difference in the periods of two pendulums based on differences in their lengths. Traditional mathematical models can be used for this task, but at the cost of specifying many numerical parameters. For many tasks such parameters are unknown. Commonsense reasoning is one such task – when reasoning about a situation expressed in a diagram [2] or via natural language, we must rely on qualitative models, since that is the level of information that is available. Here is an example from the QuaRel [3] dataset of comparative analysis problems:

> “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”

Since the distance of an episode of rolling is qualitatively inversely proportional to the friction of the floor, rolling further on the wood floor implies that (A) is the correct choice. Here the qualitative reasoning is straightforward, the complexities lie in comprehending the problem sufficiently well to enable the qualitative model to be formulated. That is what we focus on in this paper.

Natural language understanding remains a fertile source of open problems. Broadly speaking, there are two approaches to NLU today. The first is to engineer systems via machine learning, by gathering massive amounts of data, often annotated, for training ML systems, often using neural networks. With enough data and craft in dataset construction and in the training process, such systems can produce surprisingly good results on ML datasets and in some real-world applications (e.g. speech recognition, machine translation). However, such models are generally uninspectable (a problem made more concerning by their susceptibility to adversarial attacks, e.g. [4][5]), they require massive amounts of data, and they do not provide the kinds of causal explanations that traditional QR provides. The second approach is to use combinations of hand-engineering and relational learning. People do not learn language from scratch for each new task that they do. Instead, they adapt what they already know to new problems. Thus, a larger starting point of knowledge and skill enables systems to learn with higher data-efficiency. The use of expressive relational representations, especially qualitative representations, provides robustness and explainability.

Approaches that involve constructing explicit conceptual structures that capture important knowledge implied by language are what we call *high-precision understanding*. We have used this approach on many tasks now [6], including extracting useful qualitative knowledge from texts [7][8][9][10]. We believe that this approach not only yields more interpretable results, but also in the long run will be more effective in real-world problems, where distractions abound, unlike today’s carefully crafted ML datasets. Here we describe how we are applying these ideas to solving comparative analysis problems expressed in natural language text.

We begin by summarizing prior work on comparative analysis and the QuaRel dataset. We then discuss our approach, including our layered approach to NLU which enables the same language system to be tuned for many purposes, and analogical Q/A training. Then we discuss our work on progress on applying these ideas to QuaRel, along with some pilot results on the quantity identification subtask. We end with conclusions and future work.

2 Comparative Analysis and QuaRel

Two techniques have been developed to solve comparative analysis problems. The first technique is *differential qualitative analysis* (DQA) [1], which uses a set of rules to compute relative values across the descriptions of two systems, based on assumed differences between them. For example, the duration rule says that if a rate is lower in one system versus another, then the time required

[1] Computer Science department, Northwestern University, USA, email: mvcrouse@u.northwestern.edu, forbus@northwestern.edu

to reach a limit point will be longer in that system. In [1] the problem of aligning the two systems to be compared was simplified by only considering changes in parameters, but as [2] showed, the same techniques can be generalized by using analogical mappings to automatically align two systems under analysis.

The second technique is *exaggeration* [1], which uses qualitative simulation with extreme values substituted to reason about perturbations. Given a proposed perturbation to a system, exaggeration first transforms the problem with the perturbation being an extreme value, i.e. infinite if increased, zero if decreased. For example, if asked how a car would roll on a carpet if the carpet had more resistance, the reformulated model would have the resistance of the carpet being infinite, and thus the rolling speed would decrease to zero. This result is then rescaled, to respond that there would be a decrease in speed.

In our approach to QuaRel, we assume DQA with analogical mappings between situation descriptions constructed from language. Key to such descriptions are quantities identified by language, which is why we have focused on them first.

2.1 QuaRel

QuaRel uses the term *world* to denote an entity in a situation that is being compared, with each problem including two worlds. Worlds can be different entities (e.g. a rough ball versus a smooth ball) or the same entity at two different points in time. In the prior example of a car rolling on wood versus the same car rolling on carpet, the worlds would be the two rolling events, with the target comparative statement as qrel(friction, lower, world1, world2) that would be considered in conjunction with the domain theory to infer qrel(distance, higher, world1, world2). We implement worlds via Cyc-style microtheories, populated with facts extracted from language. The system defines three microtheories per question, one for each of the worlds and a microtheory for the problem as a whole, which inherits from both the world microtheories. In our system, microtheories can be treated as cases for analogical reasoning, and hence alignments can be constructed to support DQA. The set of facts in world microtheories includes the specification of which quantities apply (e.g., (hasQuantity world1-entity ((QPQuantityFn Friction) world1-entity))) and information like values (e.g., (valueOf ((QPQuantityFn Friction) world2-entity) (HighAmountFn Friction)). The general microtheory contains facts that draw direct comparisons between the entities of both worlds (e.g., (qLessThan ((QPQuantityFn Friction) world1-entity) ((QPQuantityFn Friction) world2-entity)))..

The QuaRel dataset provides a naïve domain theory[2] regarding the relationships between quantities, posed as abstract qualitative proportionalities. For example, they write q-(speed, friction) to represent that if friction goes up, speed goes down. Note that this description does not provide any scoping as to the types of entities involved. This is a common feature of informal explanations, and one of the challenges of working with natural language is flexible reasoning with such descriptions. Prior to reasoning, these descriptions would be automatically translated into the qualitative proportionalities used in NextKB and added to the world-level microtheories. For instance, with q-(speed, friction), this would mean adding the statement (qprop- ((QPQuantityFn Speed) world1-entity) ((QPQuantityFn Friction) world1-entity)) to the microtheory for the first world (and a similar statement for the second world).

QuaRel categorizes questions as either *relative value* questions or *absolute value* questions. Relative value questions have as their answer a comparative, e.g. in the car example, that the carpet has more resistance than the floor. These questions are given logical forms expressing this as (qGreaterThan ((QPQuantityFn Friction) world1-entity) ((QPQuantityFn Friction) world2-entity)). Absolute value questions still involve comparisons, but implicitly so, by using symbolic qualitative values. For example, the question “Does a bar stool slide faster along the bar surface with decorative raised bumps or the smooth wooden floor?” expresses that a bar surface with bumps has a high amount of friction (i.e., (valueOf ((QPQuantityFn Friction) bar-stool) (HighAmountFn Friction))) while a smooth bar surface has a low amount of friction.

2.2 Prior QuaRel Approaches

The original QuaRel work [3] provided two neural-based semantic parsing models that followed an encoder-decoder framework for generating logical forms. Given a question and answer option, the concatenation of the question and answer would be fed to an LSTM encoder which would produce a vector-space representation of the input text. A subsequent decoder architecture would take as input the vector representation and sequentially decode production rules from a formal grammar to build up an abstract syntax tree that would be considered the logical form. Notably, their method is completely neural, generating a logical form for both the question and answer simultaneously (i.e., no qualitative reasoner is used to generate the answer from the logical form of the question). Subsequent efforts on QuaRel have instead focused on only the multiple-choice portion of the dataset. For instance, [12] proposed translating the logical forms to text such that a BERT-based [13] textual entailment model could be used to improve multiple-choice performance. Similarly, [14] used logic-based rules to extend the training data for QuaRel and enhance a RoBERTa-based [15] multiple-choice selection model.

3 Background

We use qualitative process theory [16] for qualitative representations and reasoning because its notion of physical process and constructs have already been mapped to natural language [7][8]. Natural language processing is performed via the Companion Natural Language Understanding system (CNLU) [17]. The knowledge base for Companions and CNLU is NextKB[3], which uses representations that integrate FrameNet [18] and OpenCyc [19], as well as a broad English lexicon. We next describe enough about CNLU to understand the rest of the paper, and the idea of analogical Q/A training, which we also build upon.

3.1 Natural Language Understanding

CNLU uses multiple layers of representation to analyze language. Syntactic analysis uses Allen’s parser [20], which produces a full syntactic analysis of every sentence. Higher-level phenomena, such as counterfactuals and logical quantification, are handled by a semantic interpreter based on Discourse Representation Theory [21].

[2] https://ai2-website.s3.amazonaws.com/publications/QuaRel-Supplementary-AAAI2019.pdf

[3] http://www.qrg.northwestern.edu/nextkb/index.html, available via a Creative Commons Attribution Only license.

```
(queryCaseFor
  (and (hasQuantity world1-entity ((QPQuantityFn Strength) world1-entity))
         (hasQuantity world2-entity ((QPQuantityFn Strength) world2-entity))
  (and (comparer weak294962 world1-entity)
         (isa weak294962 ComparisonEvent)
         (comparisonQtype weak294962 Strength)
         (comparee weak294962 world2-entity)))
```

Figure 1. A query case generated for the fill-in-the-blank question, "The small child was much weaker than the adult and they _____."

Language is inherently ambiguous, and context is needed to make sense of it. CNLU represents ambiguities explicitly as *choice sets* in its analysis. Each choice set represents either a set of alternate word senses or a syntactic ambiguity. Logical constraints express the relationships between choices. For example, only one word sense for each word can be chosen in a consistent interpretation. Similarly, some word sense choices imply specific syntactic choices, and vice-versa. CNLU uses abduction to construct interpretations, based on task-specific information. This has been done via rules that detect narrative functions [17] based on task constraints. Another method, which is what we are using here, is to use *analogical Q/A training* to learn query cases for driving interpretation. We discuss this next.

3.2 Analogical Q/A Training

By design, the outputs of the semantic parser introduced in Section 3.1 are task agnostic. That is, the semantic forms it produces should be considered intermediate forms that require adaptation before being passed to task-specific reasoners (e.g., a qualitative reasoner) needed to solve a given problem. In analogical Q/A training, this adaptation is performed by analogical reasoning over *query cases*. Query cases (QCs) are rule-like constructs that treat semantic choices as antecedents and task specific logical forms as consequents. To apply a query case, the semantics for a question are aligned to the query case's antecedents via analogy, which produces an instantiation of its consequent logical form with the entities of the question at hand. Previous work has used this approach for factoid question-answering [22], process identification [23], state change prediction [24], and question-answering in a kiosk [25].

The final version of our approach will generate query cases that tie the semantic choices of each scenario to logical forms that can be passed into an off-the-shelf qualitative reasoner that performs differential qualitative analysis (DQA). With QuaRel, the challenge is not in the actual reasoning required to solve each question. Instead, the difficulty lies in understanding which quantities and relationships are relevant in the provided scenario to derive the correct answers. Though the logical forms needed for this domain are quite simple (the original QuaRel work used only simple rule-based inference as its form of reasoning), we believe that the interface we build between natural language and more powerful qualitative modeling techniques will lay the foundation for more complex reasoning over scenarios posed in natural language.

4 Our Approach

The objective of our approach is to learn query cases that can map from the natural language forms of the questions to the facts needed for DQA to infer the correct conclusions. The work presented here is still in progress, and we have yet to produce a system that can handle the full extent of the varied language found in the QuaRel dataset. Thus, our focus in this paper is to describe our approach as it has been applied to an important subproblem, determining the relevant quantities for the focus entities of a particular question. Our approach learns query cases of the form found in Figure 1. In the figure, the first argument to queryCaseFor is the consequent expression (in this case, a conjunction of quantity predictions for each of the world entities), while the second argument is the set of semantic choices produced by CNLU that was determined to be sufficient for inferring the consequent expression. In the example this query case was learned from, it was determined that a comparative relation involving Strength was sufficient for concluding that Strength was the quantity of interest.

4.1 Inducing Query Cases

Consider the question shown in Figure 1. We will refer to the question as *Q* and to its consequent logical form as *L*. We pair with *Q* a set of base statements $S_Q = \{ s_1, ..., s_i \}$ and a set of pairwise nogood constraints $N_Q = \{ n_1, ..., n_k \}$ between pairs of elements in S_Q. The elements of S_Q include the semantic choices produced for *Q* (i.e., the set of outputs generated by CNLU) as well statements indicating the root forms of words present in *Q*. The elements of N_Q are then the nogood constraints asserted between semantic choices (e.g., two semantic choices that represent alternative word senses).

Figure 2 shows a subset of the semantic parse produced for *Q*. S_Q includes semantic choices from this parse that express possible meanings of the words and phrases in the sentence. For instance, in our example question "weaker" is transformed into a set of semantic choices, two of which are (comparisonQtype weak294962 Strength) and (comparisonQtype weak294962 Effectiveness). These express different interpretations of the quantity being compared: "weaker" in the sense of strength versus "weaker" in the sense of effectiveness. The set of nogoods N_Q includes a pairwise constraint between these two quantity statements that restricts them from both being true simultaneously. An example of a word-level statement added to S_Q is (wordInSentence train-question-1741 weak). These are not alternative semantic choices, and are thus excluded from being a part of any nogood in N_Q.

```
"child"
   - (isa world1 HumanChild)
"small child"
   - (isa world1 (SmallFn HumanChild))
          ...          ...          ...
"weaker"
   - (and (isa weak294962 ComparisonEvent)
            (comparisonQtype weak294962 Strength) ...)
   - (and (isa weak294962 ComparisonEvent)
            (comparisonQtype weak294962 Effectiveness) ...)
          ...          ...          ...
"adult"
   - (isa world2 HumanAdult)
   - (isa world2 AdultAnimal)
```

Figure 2. Choices from the semantic parse of "The small child was much weaker than the adult", where the worlds being compared are the "child" and "adult"

Given a set of positive examples *Pos* (i.e., the set of all questions for which the consequent *L* is the target logical form) and negative examples *Neg* (i.e., the set of all questions for which *L* is *not* the target logical form), our approach builds a query case *QC* incrementally (with *QC* = {} initially). At each step, it selects an element s_i from S_Q to add to *QC*, i.e., $QC \cup \{ s_i \}$. We next describe how our approach picks an element to add at each step.

The primary determination of which elements from S_Q to add to *QC* is given by the information gain heuristic of FOIL [26]. Let E_1 and E_2 be sets of expressions and let N_2 be a conjunction of pairwise nogoods between the elements of E_2 (i.e., a set of mutual exclusivity constraints between elements of E_2). As E_2 and N_2 are considered an inseparable pair, we define the pair $p = (E_2, N_2)$. E_1 is said to *cover* the pair *p* if there exists a one-to-one substitution θ between the entities of E_1 and E_2 such that the following holds

$$cov(E_1, p) = (\theta E_1 \subseteq E_2) \land (\theta E_1 \land N_2 \vDash \top)$$

Informally, this means that there exists a substitution that can be applied to the set of expressions E_1 such that they can be found within E_2 and the subset of E_2 found is non-conflicting. We write that E_1 covers *p* rather than the other way around because E_1 is an abstraction that will be used to draw inferences from multiple other sets of expressions. We write the coverage score of a set of expressions to be *E*

$$E^+ = \{ p \in Pos : \ cov(E,p) \}$$

$$E^- = \{ p \in Neg : \ cov(E,p) \}$$

$$cs(E) = -\log_2 \frac{|E^+|}{|E^+| + |E^-|}$$

With these definitions, we can define the value of adding a statement s_i to *QC* as

$$gain(s_i, QC) = |QC^+| * (cs(QC \cup \{ s_i \}) - cs(QC))$$

which can be thought of as a coverage-weighted information gain heuristic. At each iteration, the element from S_Q maximizing the *gain* value is added to the query case.

Intuitively, this can be viewed as adding statements that maximize the number of positive examples with matching constituent statements while minimizing the number of negative examples with matching constituent statements. In our example, the semantic choice (isa world1 HumanChild), while a seemingly relevant choice, is actually quite useless. Such a statement appears in 5 questions regarding Strength and 22 questions involving other quantities (e.g., "A child slips more easily on ice ...", which involves Friction). Alternatively, the two choices for "weaker", being (comparisonQtype weak294962 Strength) and (comparisonQtype weak294962 Effectiveness), both appear in 17 questions regarding Strength and only 3 questions involving other quantities (e.g., "Earth has stronger gravity than Mars because ...", which involves Gravity). In this case, both of these statements have the same gain score because they cover the same numbers of positive and negative examples. To break the tie, we pick the statement that is most closely related to the target logical form *L*.

Our approach chooses between elements of S_Q with identical *gain* scores based on their relatedness to the target logical form *L*. The relatedness between two expressions is determined as in [24], using a score that measures how closely the conceptual entities in each expression are linked in the knowledge base. Formally, we define the knowledge base as a graph *G* with conceptual entities as nodes and structural relations (i.e., facts with a predicate that is a structural relation, such as resultIsa or genlPreds) as edges, the relatedness score between two concepts ce_1 and ce_1 is given by

$$rs(ce_1, ce_2) = \sum_{p \in \Psi(G, ce_1, ce_2)} \frac{\prod_{ce_k \in p} \deg(G, ce_k)^{-1}}{|\Psi(G, ce_1, ce_2)|}$$

where Ψ is a procedure that takes a graph and two conceptual entities and returns all paths connecting the entities in the given graph, and deg is a function that takes a graph and a conceptual entity and returns the out degree of the conceptual entity in the graph. The product of inverse out degree gives lower values to more common paths (i.e., paths that connect through more ubiquitous conceptual entities). The relatedness score between two expressions is then the sum of relatedness scores between each of their entities.

Relatedness leads to the preference of query cases connected through background knowledge to *L*. For our example, the statement (comparisonQtype weak294962 Strength) has a strong connection to the target logical form because they both share the conceptual entity Strength. Conversely, there is not a strong connection between the target logical form and Effectiveness, which thus means that the choice (comparisonQtype weak294962 Effectiveness) is dispreferred.

Choices from S_Q are added to *QC* until no elements from *Neg* are covered. The result is a set of expressions that can be interpreted as the antecedents to a query case (as seen in Figure 1). Additionally, we store with each query case its particular coverage of both positive and negative questions. This gives our approach a rough estimate of quality / confidence in the query case, as query cases that cover a substantial number of positive questions (and few negative questions) are more likely to be useful for answering questions than those that only cover a single positive question.

5 Pilot Experiments

Initially we have focused on extracting quantity information from problems, since without high-accuracy quantity identification, subsequent stages of understanding are likely to be very noisy. For each question, our approach first parses the question with our semantic parser to produce a set of semantic choices and constraints between said choices. Then, using the original query case framework from [22], our approach determines which quantity type is most likely applicable for the target entities of the question. As our approach can generate multiple logical forms (i.e., it can generate quantity predictions for any entity in the text, not just entities specified as the focus of the question), we filter out all predictions not relating to the entities of interest.

The training, development, and test sets for QuaRel have 1941, 278, and 552 questions, respectively. Table 1 presents our results for quantity prediction in terms of overall accuracy. On this task, random guessing would yield an accuracy of 5%. Our approach performs well above that simple baseline, however there is clear room for improvement.

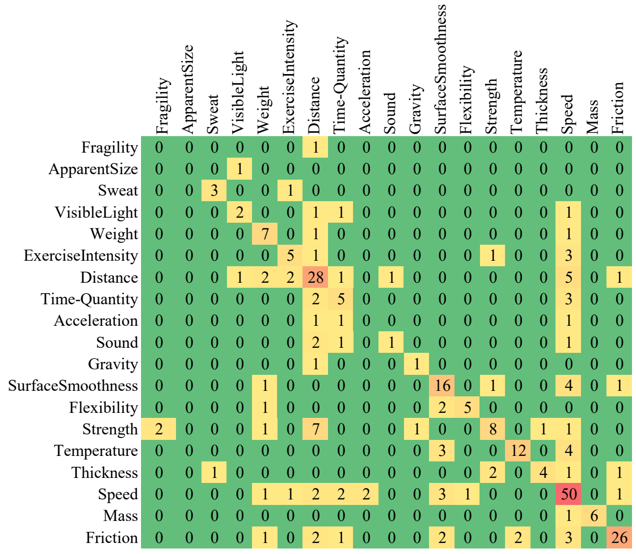
Table 1. Results for quantity prediction on the development and test sets of QuaRel

Split	*Accuracy*
Dev Set	64.4%
Test Set	59.8%

To better determine the quality of our learned query cases, we performed an error analysis on the development set where we measured which quantities tended to be confused with which other quantities. Each cell in the matrix shown in Figure 3 gives the number of times a quantity from the row was confused the quantity from the column (with the diagonal of the matrix showing correct quantity predictions). Inspecting the errors made by our approach, we can see that it tended to learn effective query cases for predicting the amount of sweat (3 / 4 predictions correct), flexibility (5 / 6 predictions correct), and friction (26 / 30 predictions correct). Conversely, it was much less effective at predicting distance (28 / 49 predictions correct), weight (7 / 14 predictions correct), time (5 / 12 predictions correct), and speed (50 / 79 predictions correct). Interestingly, out of the 99 errors, 65 of them involved mixing quantities that were in a direct causal relationship with one another (i.e., either directly influencing or being directly influenced). This provides one possible direction for improving our approach, namely, that the training procedure should try to more strongly discriminate between related quantities.

6 Conclusions

Comparative analysis is an important form of qualitative reasoning. The QuaRel dataset provides a great opportunity to explore how to do qualitative reasoning from natural language information. We have argued that a high-precision approach, in which qualitative representations play a key role, should provide more robust comprehension of such problems, compared to machine learning systems whose distributed representations provide less precision. Our initial results on quantity identification suggest that this approach is promising.



	Fragility	ApparentSize	Sweat	VisibleLight	Weight	ExerciseIntensity	Distance	Time-Quantity	Acceleration	Sound	Gravity	SurfaceSmoothness	Flexibility	Strength	Temperature	Thickness	Speed	Mass	Friction
Fragility	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
ApparentSize	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sweat	0	0	3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
VisibleLight	0	0	0	2	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0
Weight	0	0	0	0	7	0	1	0	0	0	0	0	0	0	0	0	1	0	0
ExerciseIntensity	0	0	0	0	0	5	1	0	0	0	0	0	0	1	0	0	3	0	0
Distance	0	0	0	1	2	2	28	1	0	1	0	0	0	0	0	0	5	0	1
Time-Quantity	0	0	0	0	0	0	2	5	0	0	0	0	0	0	0	0	3	0	0
Acceleration	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0
Sound	0	0	0	0	0	0	2	1	0	1	0	0	0	0	0	0	1	0	0
Gravity	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
SurfaceSmoothness	0	0	0	0	1	0	0	0	0	0	0	16	0	1	0	0	4	0	1
Flexibility	0	0	0	0	1	0	0	0	0	0	0	2	5	0	0	0	0	0	0
Strength	2	0	0	0	1	0	7	0	0	0	1	0	0	8	0	1	1	0	0
Temperature	0	0	0	0	0	0	0	0	0	0	0	3	0	0	12	0	4	0	0
Thickness	0	0	1	0	0	0	0	0	0	0	0	0	0	2	0	4	1	0	1
Speed	0	0	0	0	1	1	2	2	2	0	0	3	1	0	0	0	50	0	1
Mass	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	6	0
Friction	0	0	0	0	1	0	2	1	0	0	0	2	0	0	2	0	3	0	26

Figure 3. Quantity confusion heatmap

Our next step is to extend our techniques to handle the entire problem, including extracting the two descriptions to be compared, the relevant ordinal relationships, setting up the DQA analysis, and selecting the correct answer. We plan to adapt techniques for analogical dialogue act detection [27] to extract the two descriptions to be compared. A small amount of hand-extension of language knowledge has been done to handle comparatives that bundle quantities (e.g. "further" implies an ordinal comparison involving distance), with how those arguments are plugged into the semantics being learned by analogical Q/A training. We plan on selecting answers by interpreting them in terms of internal representations and matching on the internal representations, rather than using language-level operations, as is common in machine learning systems. In that way our technique will be usable on questions that require answers to be generated, and thereby likely to be far more robust than ML-based language model approaches. A preliminary version of this system has been implemented and achieves 34.2% in answer generation accuracy for the development set.

ACKNOWLEDGEMENTS

We thank Constantine Nakos and Danilo Ribeiro for their helpful comments on this manuscript. This research was supported by the Machine Learning, Reasoning, and Intelligence Program of the US Office of Naval Research.

REFERENCES

[1] Daniel S. Weld. 1990. Theories of comparative analysis. MIT Press, Cambridge, MA, USA.

[2] Klenk, M., Forbus, K., Tomai, E., Kim, H., and Kyckelhahn, B. (2005). Solving Everyday Physical Reasoning Problems by Analogy using Sketches. *Proceedings of AAAI 2005*.

[3] Tafjord, Oyvind, Peter Clark, Matt Gardner, Wen-tau Yih, and Ashish Sabharwal. "Quarel: A dataset and models for answering questions about qualitative relationships." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 7063-7071. 2019.

[4] Jia, Robin, and Percy Liang. "Adversarial Examples for Evaluating Reading Comprehension Systems." *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017.

[5] Marcus, Gary. "Deep learning: A critical appraisal." *arXiv preprint arXiv:1801.00631* (2018).

[6] Forbus, Kenneth D., and Thomas Hinrich. "Analogy and relational representations in the companion cognitive architecture." *AI Magazine* 38.4 (2017): 34-42.

[7] McFate, C.J., Forbus, K. and Hinrichs, T. (2014). Using Narrative Function to Extract Qualitative Information from Natural Language Texts. Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, Québec City, Québec, Canada.

[8] McFate, C., and Forbus, K. (2016). An Analysis of Frame Semantics of Continuous Processes. Proceedings of the 38th Annual Meeting of the Cognitive Science Society, Philadelphia, PA, August.

[9] McFate, C., and Forbus, K. (2016). Scaling up Linguistic Processing of Qualitative Process Interpretation. Proceedings of the Fourth Annual Conference on Advances in Cognitive Systems. Evanston, IL.

[10] Crouse, M., McFate, C.J., and Forbus, K. (2018). Learning to Build Qualitative Scenario Models from Natural Language. Proceedings of QR 2018, Stockholm.

[11] Hinrichs, T., and K. Forbus. "Toward higher-order qualitative representations." *Proceedings of the 26th International Workshop on Qualitative Reasoning*. 2012.

[12] Mitra, Arindam, Chitta Baral, Aurgho Bhattacharjee, and Ishan Shrivastava. "A Generate-Validate Approach to Answering Questions about Qualitative Relationships." *arXiv preprint arXiv:1908.03645* (2019).

[13] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[14] Asai, Akari, and Hannaneh Hajishirzi. "Logic-Guided Data Augmentation and Regularization for Consistent Question Answering." *arXiv preprint arXiv:2004.10157* (2020).

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

[16] Forbus, K. (1984). Qualitative process theory. Artificial Intelligence, 24, 85-168.

[17] Tomai, Emmett, and Kenneth D. Forbus. "EA NLU: Practical language understanding for cognitive modeling." In *Twenty-Second International FLAIRS Conference*. 2009.

[18] Baker, C.F., Fillmore, C.J. and Lowe, J.B., 1998, August. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1* (pp. 86-90). Association for Computational Linguistics.

[19] Lenat, D.B., 1995. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, *38*(11), pp.33-38.

[20] Allen, J. (1994) Natural Language Understanding, Benjamin Cummings.

[21] Kamp, H. and Reyle, U., 2013. From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory (Vol. 42). Springer Science & Business Media.

[22] Crouse, M., MFate, C.J., and Forbus, K.D. (2018). Learning from Unannotated QA Pairs to Analogically Disambiguate and Answer Questions. Proceedings of AAAI 2018.

[23] Crouse, M., McFate, C.J., & Forbus, K. (2018). Learning to Build Qualitative Scenario Models from Natural Language. Proceedings of QR 2018, Stockholm.

[24] Ribeiro, D., Hinrichs, T., Crouse, M., Forbus, K., Chang, M. and Witbrock, M., 2013. Predicting State Changes in Procedural Text using Analogical Question Answering.

[25] Wilson, J.R., Chen, K., Crouse, M., Nakos, C., Ribeiro, D.N., Rabkina, I. and Forbus, K.D., 2019. Analogical Question Answering in a Multimodal Information Kiosk.

[26] Quinlan, J.R. and Cameron-Jones, R.M., 1993, April. FOIL: A midterm report. In *European conference on machine learning* (pp. 1-20). Springer, Berlin, Heidelberg.

[27] Barbella, D. and Forbus, K. (2011). Analogical Dialogue Acts: Supporting Learning by Reading Analogies in Instructional Texts. *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI 2011),* San Francisco, CA.