CLUTRR: A Diagnostic Benchmark for Inductive Reasoning from Text

Koustuv Sinha 1,3,4, Shagun Sodhani 2,3, Jin Dong 1,3, Joelle Pineau 1,3,4 and William L. Hamilton 1,3,4
1 School of Computer Science, McGill University, Canada
2 Université de Montréal, Canada
3 Montreal Institute of Learning Algorithms (Mila), Canada
4 Facebook AI Research (FAIR), Montreal, Canada
{koustuv.sinha, sshagunsodhani, jin.dong, jpineau, wlh}
@{mail.mcgill.ca, gmail.com, mail.mcgill.ca, cs.mcgill.ca, cs.mcgill.ca}

Abstract

The recent success of natural language understanding (NLU) systems has been troubled by results highlighting the failure of these models to generalize in a systematic and robust way. In this work, we introduce a diagnostic benchmark suite, named CLUTRR, to clarify some key issues related to the robustness and systematicity of NLU systems. Motivated by classic work on inductive logic programming, CLUTRR requires that an NLU system infer kinship relations between characters in short stories. Successful performance on this task requires both extracting relationships between entities, as well as inferring the logical rules governing these relationships. CLUTRR allows us to precisely measure a model’s ability for systematic generalization by evaluating on held-out combinations of logical rules, and it allows us to evaluate a model’s robustness by adding curated noise facts. Our empirical results highlight a substantial performance gap between state-of-the-art NLU models (e.g., BERT and MAC) and a graph neural network model that works directly with symbolic inputs—with the graph-based model exhibiting both stronger generalization and greater robustness.

1 Introduction

Natural language understanding (NLU) systems have been extremely successful at reading comprehension tasks, such as question answering (QA) and natural language inference (NLI). An array of existing datasets are available for these tasks. This includes datasets that test a system’s ability to extract factual answers from text (Rajpurkar et al., 2016; Nguyen et al., 2016; Trischler et al., 2016; Mostafazadeh et al., 2016; Su et al., 2016), as well as datasets that emphasize commonsense inference, such as entailment between sentences (Bowman et al., 2015; Williams et al., 2018).

Figure 1: CLUTRR inductive reasoning task.

However, there are growing concerns regarding the ability of NLU systems—and neural networks more generally—to generalize in a systematic and robust way (Bahdanau et al., 2019; Lake and Baroni, 2018; Johnson et al., 2017). For instance, recent work has highlighted the brittleness of NLU systems to adversarial examples (Jia and Liang, 2017), as well as the fact that NLU models tend to exploit statistical artifacts in datasets, rather than exhibiting true reasoning and generalization capabilities (Gururangan et al., 2018; Kaushik and Lipton, 2018). These findings have also dovetailed with the recent dominance of large pre-trained language models, such as BERT, on NLU benchmarks (Devlin et al., 2018; Peters et al., 2018), which suggest that the primary difficulty in these datasets is incorporating the statistics of the natural language, rather than reasoning.

An important challenge is thus to develop NLU benchmarks that can precisely test a model’s capability for robust and systematic generalization. Ideally, we want language understanding systems that can not only answer questions and draw inferences from text, but that can also do so in a systematic, logical, and robust way. While such reasoning capabilities are certainly required for many existing NLU tasks, most datasets combine several challenges of language understanding into one, such as co-reference/entity resolution, incorporating world knowledge, and semantic parsing—making it difficult to isolate and diagnose a model’s capabilities for systematic generalization and robustness.
Our work. Inspired by the classic AI challenge of inductive logic programming (Quinlan, 1990)—as well as the recently developed CLEVR dataset for visual reasoning (Johnson et al., 2017)—we propose a semi-synthetic benchmark designed to explicitly test an NLU model’s ability for systematic and robust logical generalization.

Our benchmark suite—termed CLUTRR (Compositional Language Understanding and Text-based Relational Reasoning)—contains a large set of semi-synthetic stories involving hypothetical families. Given a story, the goal is to infer the relationship between two family members, whose relationship is not explicitly mentioned (Figure 1). To solve this task, a learning agent must extract the relationships mentioned in the text, induce the logical rules governing the kinship relationships (e.g., the transitivity of the sibling relation), and use these rules to infer the relationship between a given pair of entities. Crucially, the CLUTRR benchmark allows us to test a learning agent’s ability for systematic generalization by testing on stories that contain unseen combinations of logical rules. CLUTRR also allows us to precisely test for the various forms of model robustness by adding different kinds of superfluous noise facts to the stories.

We compare the performance of several state-of-the-art NLU systems on this task—including Relation Networks (Santoro et al., 2017), Compositional Attention Networks (Hudson and Manning, 2018) and BERT (Devlin et al., 2018). We find that the generalization ability of these NLU systems is substantially below that of a Graph Attention Network (Veličković et al., 2018), which is given direct access to symbolic representations of the stories. Moreover, we find that the robustness of the NLU systems generally does not improve by training on noisy data, whereas the GAT model is able to effectively learn robust reasoning strategies by training on noisy examples. Both of these results highlight important open challenges for closing the gap between machine reasoning models that work with unstructured text and models that are given access to more structured input.

2 Related Work

We draw inspiration from the classic work on inductive logic programming (ILP), a long line of reading comprehension benchmarks in NLP, as well as work combining language and knowledge graphs.

Reading comprehension benchmarks. Many datasets have been proposed to test the reading comprehension ability of NLP systems. This includes the SQuAD (Rajpurkar et al., 2016), NewsQA (Trischler et al., 2016), and MCTest (Richardson et al., 2013) benchmarks that focus on factual questions; the SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) benchmarks for sentence understanding; and the bAbI tasks (Weston et al., 2015), to name a few. Our primary contribution to this line of work is the development of a carefully designed diagnostic benchmark to evaluate model robustness and systematic generalization in the context of NLU.

Question-answering with knowledge graphs. Our work is also related to the domain of question answering and reasoning in knowledge graphs (Das et al., 2018; Xiong et al., 2018; Hamilton et al., 2018; Wang et al., 2018; Xiong et al., 2017; Welbl et al., 2018; Kartsaklis et al., 2018), where either the model is provided with a knowledge graph to perform inference over or where the model must infer a knowledge graph from the text itself. However, unlike previous benchmarks in this domain—which are generally transductive and focus on leveraging and extracting knowledge graphs as a source of background knowledge about a fixed set of entities—CLUTRR requires inductive logical reasoning, where every example requires reasoning over a new set of previously unseen entities.

3 Benchmark Design

In order to design an NLU benchmark that explicitly tests inductive reasoning and systematic generalization, we build upon the classic ILP task of inferring family (i.e., kinship) relations (Hinton et al., 1986; Muggleton, 1991; Lavrac and Dzeroski, 1994; Kok and Domingos, 2007; Rocktaschel and Riedel, 2017). For example, given the facts that “Alice is Bob’s mother” and “Jim is Alice’s father”, one can infer with reasonable certainty that “Jim is Bob’s grandfather.” While this example may appear trivial, it is a challenging task to design models that can learn from data to induce the logical rules necessary to make such inferences, and it is even more challenging to design models that can systematically generalize by composing these induced rules.

Inspired by this classic task of logical induction and reasoning, the CLUTRR benchmark requires an NLU system to infer and reason about kinship
relations by reading short stories. Requiring that the models learn directly from natural language makes this task much more challenging than the purely symbolic ILP setting. However, we leverage insights from traditional ILP to generate these stories in a semi-synthetic manner, providing precise control over the complexity of the reasoning required to solve the task.

In its entirety, the CLUTRR benchmark suite allows researchers to generate diverse semi-synthetic short stories to test different aspects of inductive reasoning capabilities. We publicly release the entire benchmark suite, including code to generate the semi-synthetic examples, the specific datasets that we introduce here, and the different baselines that we compare with.¹

3.2 Story generation

The short stories in CLUTRR are essentially narrativized renderings of a set of logical facts. In this section, we describe how we sample the logical facts that make up a story by generating random kinship graphs and using backward chaining to produce logical reasoning chains. The conversion from logical facts to natural language narratives is then described in Section 3.3.

Terminology and background. Following standard practice in formal semantics, we use the term atom to refer to a predicate symbol and a list of terms, such as \( \text{grandfatherOf}(X,Y) \), where the predicate \text{grandfatherOf} denotes the relation between the two variables, \( X \) and \( Y \). We restrict the predicates to have an arity of 2, i.e., binary predicates. A logical rule in this setting is of the form \( H \vdash B \), where \( B \) is the body of the rule, i.e., a conjunction of two atoms \((\alpha_1, \alpha_2)\) and \( H \) is the head, i.e., a single atom \((\alpha)\) that can be viewed as the goal or query. For instance, given a knowledge base (KB) \( R \) that contains the single rule \( \text{grandfatherOf}(X,Y) \vdash (\text{fatherOf}(X,Z), \text{fatherOf}(Z,Y)) \), the query \( \text{grandfatherOf}(X,Y) \) evaluates to true if and only if the body \( B = (\text{fatherOf}(X,Z), \text{fatherOf}(Z,Y)) \) is also true in a given world. A rule is called a grounded rule if all atoms in the rule are themselves grounded, i.e., all variables are replaced with constants or entities in a world. A fact is a grounded binary predicate. A clause is a conjunction of two or more atoms \((C = (H_C \vdash B_C = ((\alpha_1, ..., \alpha_n)))\) which can be built using a set of rules.

In the context of our data generation process, we distinguish between the knowledge base, \( R \), which contains a finite number of predicates and rules specifying how kinship relations in a family resolve, and a particular kinship graph \( G \), which

---

¹Benchmark suite code can be obtained from https://github.com/facebookresearch/clutrr. Generated datasets are available to view in this link.
contains a grounded set of atoms specifying the particular kinship relations that underlie a single story. In other words, $R$ contains the logical rules that govern all the generated stories in CLUTRR, while $G$ contains the grounded facts that underlie a specific story.

**Graph generation.** To generate the kinship graph $G$ underlying a particular story, we first sample a set of gendered entities and kinship relations using a stochastic generation process. This generation process contains a number of tunable parameters—such as the maximum number of children at each node, the probability of an entity being married to another entity, etc.—and is designed to produce a valid, but possibly incomplete “backbone graph”. For instance, this backbone graph generation process will specify “parent”/“child” relations between entities but does not add “grandparent” relations. After this initial generation process, we recursively apply the logical rules in $R$ to the backbone graph to produce a final graph $G$ that contains the full set of kinship relations between all the entities.

**Backward chaining.** The resulting graph $G$ provides the background knowledge for a specific story, as each edge in this graph can be treated as a grounded predicate (i.e., fact) between two entities. From this graph $G$, we sample the facts that make up the story, as well as the target fact that we seek to predict: First, we (uniformly) sample a target relation $H_C$, which is the fact that we want to predict from the story. Then, from this target relation $H_C$, we run a simple variation of the backward chaining (Gallaire and Minker, 1978) algorithm for $k$ iterations starting from $H_C$, where at each iteration we uniformly sample a subgoal to resolve and then uniformly sample a KB rule that resolves this subgoal. Crucially, unlike traditional backward chaining, we do not stop the algorithm when a proof is obtained; instead, we run for a fixed number of iterations $k$ in order to sample a set of $k$ facts $B_C$ that imply the target relation $H_C$.

**3.3 Adding natural language**

So far, we have described the process of generating a conjunctive logical clause $C = (H_C \vdash B_C)$, where $H_C = \{\alpha\}$ is the target fact (i.e., relation) we seek to predict and $B_C = \{\alpha_1, \ldots, \alpha_k\}$ is the set of supporting facts that imply the target relation. We now describe how we convert this logical representation to natural language through crowd-sourcing.

**Paraphrasing using Amazon Mechanical Turk.** The basic idea behind our approach is that we show Amazon Mechanical Turk (AMT) crowd-workers the set of facts $B_C$ corresponding to a story and ask the workers to paraphrase these facts into a narrative. Since workers are given a set of facts $B_C$ to work from, they are able to combine and split multiple facts across separate sentences and construct diverse narratives (Figure 3). Appendix 1.6 contains further details on our AMT interface (based on the ParlAI framework (Miller et al., 2017)), data collection, and the quality controls we employed.

**Reusability and composition.** One challenge for data collection via AMT is that the number of possible stories generated by CLUTRR grows combinatorially as the number of supporting facts increases, i.e., as $k = |B_C|$ grows. This combinatorial explosion for large $k$—combined with the difficulty of maintaining the quality of the crowd-sourced paraphrasing for long stories—makes it infeasible to obtain a large number of paraphrased examples for $k > 3$. To circumvent this issue and increase the flexibility of our benchmark, we reuse and compose AMT paraphrases to generate longer stories. In particular, we collected paraphrases for stories containing $k = 1, 2, 3$ supporting facts and then replaced the entities from these collected stories with placeholders in order to re-use them to generate longer semi-synthetic stories. An example of a story generated by stitching together two shorter paraphrases is provided below:

- [Frank] went to the park with his father, [Brett].
- [Frank] called his brother [Boyd] on the phone.
- He wanted to go out for some beers. [Boyd] went to the baseball game with his son [Jim].

Q: What is [Brett] and [Jim]’s relationship?

Thus, instead of simply collecting paraphrases for a fixed number of stories, we instead obtain a diverse
Table 1: Statistics of the AMT paraphrases. Jaccard word overlap is calculated within the templates of each individual clause of length \( k \).

<table>
<thead>
<tr>
<th>Number of Paraphrases</th>
<th>( k = 1 )</th>
<th>( k = 2 )</th>
<th>( k = 3 )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1,868</td>
<td>1,890</td>
<td>2,258</td>
<td>6,016</td>
</tr>
<tr>
<td>Unigrams</td>
<td>0.201</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bigrams</td>
<td>0.0385</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

collection of natural language templates that can be programmatically recombined to generate stories with various properties.

**Dataset statistics.** At the time of submission, we have collected 6,016 unique paraphrases with an average of 19 paraphrases for every possible logical clause of length \( k = 1, 2, 3 \). Table 1 contains summary statistics of the collected paraphrases. Overall, we found high linguistic diversity in the collected paraphrases. For instance, the average Jaccard overlap in unigrams between pairs paraphrases corresponding to the same logical clause was only 0.201 and only 0.0385 for bigrams. The Appendix contains further examples of the paraphrases.

**Human performance.** To get a sense of the data quality and difficulty involved in CLUTRR, we asked human annotators to solve the task for random examples of length \( k = 2, 3, \ldots, 6 \). We found that time-constrained AMT annotators performed well (i.e., > 70%) accuracy for \( k \leq 3 \) but struggled with examples involving longer stories, achieving 40-50% accuracy for \( k > 3 \). However, trained annotators with unlimited time were able to solve 100% of the examples (Appendix 1.7), highlighting the fact that this task requires attention and involved reasoning, even for humans.

3.4 Query representation and inference

**Representing the question.** The AMT paraphrasing approach described above allows us to convert the set of supporting facts \( B_C \) to a natural language story, which can be used to predict the target relation/query \( H_C \). However, instead of converting the target query, \( H_C = [a^*] \), to a natural language question, we instead opt to represent the target query as a \( K \)-way classification task, where the two entities in the target relation are provided as input and the goal is to classify the relation that holds between these two entities. This representation avoids the pitfall of revealing information about the answer in the question (Kaushik and Lipton, 2018).

**Representing entities.** When generating stories, entity names are randomly drawn from a set of 300 common gendered English names. Thus, depending on each run, the entities are never the same. This ensures that the entity names are simply placeholders and uncorrelated from the task.

3.5 Variants of CLUTRR

The modular nature of CLUTRR provides rich diagnostic capabilities for evaluating the robustness and generalization abilities of neural language understanding systems. We highlight some key diagnostic capabilities available via different variations of CLUTRR below. These diagnostic variations correspond to the concrete datasets that we generated in this work, and we describe the results on these Datasets in Section 4.

**Systematic generalization.** Most prominently, CLUTRR allows us to explicitly evaluate a model’s ability for systematic generalization. In particular, we rely on the following hold-out procedures to test systematic generalization:

- During training, we hold out a subset of the collected paraphrases, and we only use this held-out subset of paraphrases when generating the test set. Thus, to succeed on CLUTRR, an NLU system must exhibit linguistic generalization and be robust to linguistic variation at test time.
- We also hold out a subset of the logical clauses during training (for clauses of length \( k > 2 \)). In other words, during training, the model sees all logical rules but does not see all combinations of these logical rules. Thus, in addition to linguistic generalization, success on this task also requires logical generalization.
- Lastly, as a more extreme form of both logical and linguistic generalization, we consider the setting where the models are trained on stories generated from clauses of length \( \leq k \) and evaluated on stories generated from larger clauses of length \( > k \). Thus, we explicitly test the ability for models to generalize on examples that require more steps of reasoning that any example they encountered during training.

\(^3\)One should not holdout clauses from length \( k = 2 \) in order to allow models to learn the compositionality of all possible binary predicates.
Robust Reasoning. In addition to evaluating systematic generalization, the modular setup of CLUTRR also allows us to diagnose model robustness by adding noise facts to the generated narratives. Due to the controlled semi-synthetic nature of CLUTRR, we are able to provide a precise taxonomy of the kinds of noise facts that can be added (Figure 4). In order to structure this taxonomy, it is important to recall that any set of supporting facts $B_C$ generated by CLUTRR can be interpreted as a path, $p_C$, in the corresponding kinship graph $G$ (Figure 2). Based on this interpretation, we view adding noise facts from the perspective of sampling three different types of noise paths, $p_n$, from the kinship graph $G$:

- **Irrelevant facts**: We add a path $p_n$, which has exactly one shared end-point with $p_C$. In this way, this is a *distractor* path, which contains facts that are connected to one of the entities in the target relation, $H_C$, but do not provide any information that could be used to help answer the query.

- **Supporting facts**: We add a path $p_n$, whose two end-points are on the path $p_C$. The facts on this path $p_n$ are noise because they are not needed to answer the query, but they are supporting facts because they can, in principle, be used to construct alternative (longer) reasoning paths that connect the two target entities.

- **Disconnected facts**: We add paths which neither originate nor end in any entity on $p_C$. These disconnected facts involve entities and relations that are completely unrelated to the target query.

4 Experiments

We evaluate several neural language understanding systems on the proposed CLUTRR benchmark to surface the relative strengths and shortcomings of these models in the context of inductive reasoning and combinatorial generalization.4 We aim to answer the following key questions:

(Q1) How do state-of-the-art NLU models compare in terms of systematic generalization? Can these models generalize to stories with unseen combinations of logical rules?

(Q2) How does the performance of neural language understanding models compare to a graph neural network that has full access to graph structure underlying the stories?

(Q3) How robust are these models to the addition of noise facts to a given story?

4.1 Baselines

Our primary baselines are neural language understanding models that take unstructured text as input. We consider bidirectional LSTMs (Hochreiter and Schmidhuber, 1997; Cho et al., 2014) (with and without attention), as well as recently proposed models that aim to incorporate inductive biases towards relational reasoning: Relation Networks (RN) (Santoro et al., 2017) and Compositional Memory Attention Network (MAC) (Hudson and Manning, 2018). We also use the large pretrained language model, BERT (Devlin et al., 2018), as well as a modified version of BERT having a trainable LSTM encoder on top of the pre-trained BERT embeddings. All of these models (except BERT) were re-implemented in PyTorch 1.0 (Paszke et al., 2017) and adapted to work with the CLUTRR benchmark.

Since the underlying relations in the stories generated by CLUTRR inherently form a graph, we also experiment with a Graph Attention Network (GAT) (Veličković et al., 2018). Rather than taking the textual stories as input, the GAT baseline receives a structured graph representation of the facts that underlie the story.

Entity and query representations. We use the various baseline models to encode the natural language story (or graph) into a fixed-dimensional embedding. With the exception of the BERT models, we do not use pre-trained word embeddings and learn the word embeddings from scratch using end-to-end backpropagation. An important note, however, is that we perform Cloze-style anonymization (Hermann et al., 2015) of the entities (i.e., names) in the stories, where each entity name is replaced by a @entity-k placeholder, which is randomly sampled from a small, fixed pool of placeholder tokens. The embeddings for these placeholders are randomly initialized and fixed during training.5

---

4Code to reproduce all the results in this section will be released at https://github.com/facebookresearch/clutrr/.

5See Appendix 1.5 for a comparison of placeholder em-
To make a prediction about a target query given a story, we concatenate the embedding of the story (generated by the baseline model) with the embeddings of the two target entities and we feed this concatenated embedding to a 2-layer feed-forward neural network with a softmax prediction layer.

4.2 Experimental Setup

Hyperparameters. We selected hyperparameters for all models using an initial grid search on the systematic generalization task (described below). All models were trained for 100 epochs with Adam optimizer and a learning rate of 0.001. The Appendix provides details on the selected hyperparameters.

Generated datasets. For all experiments, we generated datasets with 10-15k training examples. In many experiments, we report training and testing results on stories with different clause lengths $k$. (For brevity, we use the phrase “clause length” throughout this section to refer to the value $k = |\mathcal{B}_C|$, i.e., the number of steps of reasoning that are required to predict the target query.) In all cases, the training set contains 5000 train stories per $k$ value, and, during testing, all experiments use 100 test stories per $k$ value. All experiments were run 10 times with different randomly generated stories, and means and standard errors over these 10 runs are reported. As discussed in Section 3.5, during training we holdout 20% of the paraphrases, as well as 10% of the possible logical clauses.

4.3 Results and Discussion

With our experimental setup in place, we now address the three key questions (Q1-Q3) outlined at the beginning of Section 4.

Q1: Systematic Generalization

We begin by using CLUTRR to evaluate the ability of the baseline models to perform systematic generalization (Q1). In this setting, we consider two training regimes: in the first regime, we train all models with clauses of length $k = 2, 3$, and in the second regime, we train with clauses of length $k = 2, 3, 4$. We then test the generalization of these models on test clauses of length $k = 2, ..., 10$.

Figure 5 illustrates the performance of different models on this generalization task. We observe that the GAT model is able to perform near-perfectly on the held-out logical clauses of length $k = 3$, with the BERT-LSTM being the top-performer among the text-based models but still significantly below the GAT. Not surprisingly, the performance of all models degrades monotonically as we increase the length of the test clauses, which highlights the challenge of “zero-shot” systematic generalization (Lake and Baroni, 2018; Sodhani et al., 2018). However, as expected, the performance of all models improves on their generalization performance when trained on $k = 2, 3, 4$ rather than just $k = 2, 3$ (Figure 5, right). The GAT, in particular, achieves the biggest gain by this expanded training.

Q2: The Benefit of Structure

The empirical results on systematic generalization also provide insight into how the text-based NLU systems compare against the graph-based GAT model that has full access to the logical graph structure underlying the stories (Q2). Indeed, the relatively strong performance of the GAT model (Figure 5) suggests that the language-based models fail to learn a robust mapping from the natural language narratives to the underlying logical facts.
Table 2: Testing the robustness of the various models when training and testing on stories containing various types of noise facts. The types of noise facts (supporting, irrelevant, and disconnected) are defined in Section 3.5.

<table>
<thead>
<tr>
<th>Models</th>
<th>Unstructured models (no graph)</th>
<th>Structured model (with graph)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>Testing</td>
<td>BilSTM - Attention</td>
</tr>
<tr>
<td>Clean Clean</td>
<td>0.58 ± 0.05</td>
<td>0.53 ± 0.05</td>
</tr>
<tr>
<td>Supporting</td>
<td>0.76 ± 0.03</td>
<td>0.64 ± 0.02</td>
</tr>
<tr>
<td>Irrelevant</td>
<td>0.7 ± 0.03</td>
<td>0.76 ± 0.02</td>
</tr>
<tr>
<td>Disconnected</td>
<td>0.49 ± 0.03</td>
<td>0.45 ± 0.03</td>
</tr>
<tr>
<td>Supporting</td>
<td>0.67 ± 0.04</td>
<td>0.66 ± 0.07</td>
</tr>
<tr>
<td>Irrelevant</td>
<td>0.51 ± 0.03</td>
<td>0.47 ± 0.09</td>
</tr>
<tr>
<td>Disconnected</td>
<td>0.57 ± 0.04</td>
<td>0.57 ± 0.06</td>
</tr>
<tr>
<td>Average</td>
<td>0.61 ± 0.06</td>
<td>0.59 ± 0.06</td>
</tr>
</tbody>
</table>

To further confirm this trend, we ran experiments with modified train and test splits for the text-based models, where the same set of natural language paraphrases were used to construct the narratives in both the train and test splits (see Appendix 1.3 for details). In this simplified setting, the text-based models must still learn to reason about held-out logical patterns, but the difficulty of parsing the natural language is essentially removed, as the same natural language paraphrases are used during testing and training. We found that the text-based models were competitive with the GAT model in this simplified setting (Appendix Figure 1), confirming that the poor performance of the text-based models on the main task is driven by the difficulty of parsing the unseen natural language narratives.

### Q3: Robust Reasoning

Finally, we use CLUTRR to systematically evaluate how various baseline neural language understanding systems cope with noise (Q3). In all the experiments we provide a combination of \( k = 2 \) and \( k = 3 \) length clauses in training and testing, with noise facts being added to the train and/or test set depending on the setting (Table 2). We use the different types of noise facts defined in Section 3.5.

Overall, we find that the GAT baseline outperforms the unstructured text-based models across most testing scenarios (Table 2), which showcases the benefit of a structured feature space for robust reasoning. When training on clean data and testing on noisy data, we observe two interesting trends that highlight the benefits and shortcomings of the various model classes:

1. All the text-based models excluding BERT actually perform better when testing on examples that have supporting or irrelevant facts added. This suggests that these models actually benefit from having more content related to the entities in the story. Even though this content is not strictly useful or needed for the reasoning task, it may provide some linguistic cues (e.g., about entity genders) that the models exploit. In contrast, the BERT-based models do not benefit from the inclusion of this extra content, which is perhaps due to the fact that they are already built upon a strong language model (e.g., that already adequately captures entity genders.)

2. The GAT model performs poorly when supporting facts are added but has no performance drop when disconnected facts are added. This suggests that the GAT model is sensitive to changes in the story. Even though this content is disconnected from the target entities.

Moreover, when we trained on noisy examples, we found that only the GAT model was able to consistently improve its performance (Table 2). Again, this highlights the performance gap between the unstructured text-based models and the GAT.

### 5 Conclusion

In this paper we introduced the CLUTRR benchmark suite to test the systematic generalization and inductive reasoning capabilities of NLU systems. We demonstrated the diagnostic capabilities of CLUTRR and found that existing NLU systems exhibit relatively poor robustness and systematic generalization capabilities—especially when compared to a graph neural network that works directly with symbolic input. These results highlight the gap that remains between machine reasoning models that work with unstructured text and models that are given access to more structured input. We hope that by using this benchmark suite, progress can be made in building more compositional, modular, and robust NLU systems.
6 Acknowledgements

The authors would like to thank Jack Urbanek, Stephen Roller, Adina Williams, Dzmitry Bahdanau, Prasanna Parthasarathy, Harsh Satija for useful discussions and technical help. The authors would also like to thank Abhishek Das, Carlos Eduardo Lassance, Gunshi Gupta, Milan Aggarwal, Rim Assouel, Weiping Song, and Yue Dong for feedback on the draft. The authors also like to thank the many anonymous Mechanical Turk participants for providing paraphrases, and thank Sumana Basu, Etienne Denis, Jonathan Lebensold, and Komal Teru for providing human performance measures. The authors would also like to thank Sanghyun Yoo, Jehun Jeon and Dr Young Sang Choi of Samsung Advanced Institute of Technology (SAIT) for supporting the previous workshop version of this work. The authors are grateful to Facebook AI Research (FAIR) for providing extensive compute and GPU resources and support. This research was supported by the Canada CIFAR Chairs in AI program.

References


Index

1 Appendix

1.1 Implementation details of the baseline models

Setup. We implemented all the models using the encoder-decoder architecture. The encoders are different baseline models (listed below). The encoder takes as input the given story (paragraph ) and produces the representation of the story. In all the models, the decoder is implemented as a 2-layer MLP which takes as input the concatenated representation of the story and the embedding of the entities (for which the relationship is to be predicted) and returns a softmax distribution over the relation types. We now describe the different baseline models (encoders) in detail:

LSTM (Hochreiter and Schmidhuber, 1997): The input paragraph is processed by a two-layer Bidirectional LSTM and the hidden state corresponding to the last time-step is used as the representation of the story.

LSTM+attention (Cho et al., 2014): Similar to LSTM, but instead of using just the hidden state at the last timestep, the model computes the attention-weighted mean of the hidden state at all time steps to use as the representation of the story.

Relation Networks - RN (Santoro et al., 2017): An relation module (implemented as an MLP) is used alongside the LSTM to learn pairwise relations among all the pairs of sentences. These relation representations are the output of the relational module. Our input data is prepared as a batch of sentences x words. Each sentence is fed to the LSTM, followed by a pooling (e.g. mean, max) over all hidden states of each sentence to generate the sentence embeddings. The query embeddings are no longer needed in the decoder since they have been incorporated by the relational module when learning relations between sentences.

MAC (Compositional Attention Network) (Hudson and Manning, 2018): A MAC cell is similar to RN, but it also which contains a control state \( c \) and memory state \( m \) and can iterate over the input several times. The number of iterations is a hyperparameter. Just like RN, MAC is added behind the LSTM. In each iteration, the model attends to the embeddings of the query entities to generate the current control \( c_i \). Another attention head over \( c_i \) and all hidden outputs of LSTM is used to distill the new information \( r_i \). In the end, a linear layer is used to generate the new memory \( m_i \) by combining \( r_i \) and \( m_{i-1} \). The final memory state gives the representation of the story. This model is the state-of-the-art model for the CLEVER task.

BERT (Devlin et al., 2018): We adapt BERT pretrained language model to our task. Specifically, we use two variants of BERT - the vanilla 12-layered frozen BERT with pre-trained embeddings, and BERT-LSTM, where a one-layer LSTM encoder is added on top of pretrained BERT embeddings. BERT encodes the sentences into 768-dimensional vectors. To ensure that BERT does not treat the entities as unknown tokens (and hence producing the same representation for all of them), we represent the entities with numbers in the vanilla BERT setup. In BERT-LSTM, we replace the entity embeddings by our entity embedding lookup policy (Refer Appendix 1.5). In both the cases, we use a simple two-layer MLP decoder which takes as inputs the pooled document representation and query representations and produces the softmax distribution over the relations.

Graph Attention Network (GAT) (Velićković et al., 2018): Entity (modelled as nodes in the graph) representations are learned by using the GAT Graph Neural Network with attention-based aggregation over the neighbor nodes. We modify the GAT architecture by attending over each node \( v_j \) in the neighborhood of \( v_i \) by concatenating the edge representation \( e_{i,j} \) to the representation of \( v_i \).

Relational Recurrent Network (RMC) (Santoro et al., 2018): We also implemented RMC, a recently proposed model for relational reasoning. It works like an RNN, processing words step-by-step, except that a memory matrix \( (num_{slots} \times mem_{size}) \) is added as the hidden state. The relational bias is extracted in each step by using self-attention over the concatenation of memory matrix and the word input in the current step. The final memory matrix is the representation of the story. Our implementation is based on another open source implementation.  

1.2 Implementation details of the model

The final memory state gives the representation of the story. This model is the state-of-the-art model for the CLEVER task.

BERT (Devlin et al., 2018): We adapt BERT pretrained language model to our task. Specifically, we use two variants of BERT - the vanilla 12-layered frozen BERT with pre-trained embeddings, and BERT-LSTM, where a one-layer LSTM encoder is added on top of pretrained BERT embeddings. BERT encodes the sentences into 768-dimensional vectors. To ensure that BERT does not treat the entities as unknown tokens (and hence producing the same representation for all of them), we represent the entities with numbers in the vanilla BERT setup. In BERT-LSTM, we replace the entity embeddings by our entity embedding lookup policy (Refer Appendix 1.5). In both the cases, we use a simple two-layer MLP decoder which takes as inputs the pooled document representation and query representations and produces the softmax distribution over the relations.

Graph Attention Network (GAT) (Velićković et al., 2018): Entity (modelled as nodes in the graph) representations are learned by using the GAT Graph Neural Network with attention-based aggregation over the neighbor nodes. We modify the GAT architecture by attending over each node \( v_j \) in the neighborhood of \( v_i \) by concatenating the edge representation \( e_{i,j} \) to the representation of \( v_i \).

Relational Recurrent Network (RMC) (Santoro et al., 2018): We also implemented RMC, a recently proposed model for relational reasoning. It works like an RNN, processing words step-by-step, except that a memory matrix \( (num_{slots} \times mem_{size}) \) is added as the hidden state. The relational bias is extracted in each step by using self-attention over the concatenation of memory matrix and the word input in the current step. The final memory matrix is the representation of the story. Our implementation is based on another open source implementation.  

We noticed that the performance of the model is significantly less across all the tasks by a large margin. Till the time of the submission, we could not verify whether this subpar performance is due to buggy implementation of the code or due to some unexplored hyperparameter combination. Hence we decided not to include the results corresponding to this model in the empirical evaluation. We will continue working on verifying the implementation of the model.

https://github.com/L05G/relational-rnn-pytorch
1.2 Relations and KB used in CLUTRR Benchmark

\[
\begin{align*}
grand, X, Y \rightarrow & \langle \text{child}, X, Z \rangle, \langle \text{child}, Z, Y \rangle, \\
grand, X, Y \rightarrow & \langle \text{so}, X, Z \rangle, \langle \text{grand}, Z, Y \rangle, \\
grand, X, Y \rightarrow & \langle \text{grand}, X, Z \rangle, \\
\text{sibling}, Z, Y \rangle, \\
\text{inv-grand}, X, Y \rightarrow & \langle \text{inv-child}, X, Z \rangle, \\
\text{inv-grand}, X, Y \rightarrow & \langle \text{inv-child}, Z, Y \rangle, \\
\text{inv-grand}, X, Y \rightarrow & \langle \text{sibling}, X, Z \rangle, \\
\text{inv-grand}, Z, Y \rangle, \\
\text{child}, X, Y \rightarrow & \langle \text{child}, X, Z \rangle, \\
\text{sibling}, Z, Y \rangle, \\
\text{child}, X, Y \rightarrow & \langle \so, X, Z \rangle, \\
\text{child}, Z, Y \rangle, \\
\text{inv-child}, X, Y \rightarrow & \langle \text{inv-child}, Z, Y \rangle, \\
\text{inv-child}, X, Y \rightarrow & \langle \text{sibling}, X, Z \rangle, \\
\text{inv-child}, Z, Y \rangle, \\
\text{child}, X, Y \rightarrow & \langle \text{child}, X, Z \rangle, \\
\text{inv-un}, Z, Y \rangle, \\
\text{sibling}, X, Y \rightarrow & \langle \text{inv-child}, X, Z \rangle, \\
\text{child}, Z, Y \rangle, \\
\text{sibling}, X, Y \rightarrow & \langle \text{inv-child}, X, Z \rangle, \\
\text{child}, Z, Y \rangle, \\
\text{in-law}, X, Y \rightarrow & \langle \text{child}, X, Z \rangle, \\
\text{so}, Z, Y \rangle, \\
\text{inv-in-law}, X, Y \rightarrow & \langle \so, X, Z \rangle, \\
\text{inv-child}, Z, Y \rangle, \\
\text{un}, X, Y \rightarrow & \langle \text{sibling}, X, Z \rangle, \\
\text{child}, Z, Y \rangle, \\
\text{inv-un}, X, Y \rightarrow & \langle \text{inv-child}, X, Z \rangle, \\
\text{sibling}, Z, Y \rangle, \\
\end{align*}
\]

In the CLUTRR Benchmark, the following kinship relations are used: son, father, husband, brother, grandson, grandfather, son-in-law, father-in-law, brother-in-law, uncle, nephew, daughter, mother, wife, sister, granddaughter, grandmother, daughter-in-law, mother-in-law, sister-in-law, aunt, niece.

We used a small, tractable, and logically sound KB of rules as mentioned above. We carefully select this set of deterministic rules to avoid ambiguity in the resolution. We use gender-neutral predicates and resolve the gender of the predicate in the head \( H \) of a clause \( C \) by deducing the gender of the second constant. We have two types of predicates, \emph{vertical} predicates (parent-child relations) and \emph{horizontal} predicates (sibling or significant other). We denote all the vertical predicates by its \textit{child-to-parent} relation and append the prefix \texttt{inv-} to the predicates for the corresponding \textit{parent-to-child} relation. For example, \texttt{grandfatherOf} is denoted by the gender-neutral predicate \texttt{[inv-grand, X, Y]}, where the gender is determined by the gender of \( Y \).

1.3 Effect of placeholder size and split

To analyze whether the language models fail to learn a robust mapping from natural language narratives to underlying logical facts, we re-run the generalization experiments with a reduced placeholder size (20% of the full collected placeholders) \emph{and} we keep the same placeholders for both training and testing. We observe all language-based models are now competitive with respect to GAT on both training regimes \( k = 2, 3 \) and \( k = 2, 3, 4 \). This shows the need for separating the placeholder split to effectively test systematic generalization because otherwise, current NLU systems tend to exploit the underlying language layer to arrive at the correct answer.

1.4 More evaluations on Robust Reasoning

We performed several additional experiments to analyze the effect of different training regimes in the Robust Reasoning setup (Table 3) of CLUTRR. Specifically, we want to analyze the effect on zero-shot generalization and robustness when training with different noisy data settings. We notice that the GAT model, having access to the true underlying graph of the puzzles, perform better across different testing scenarios when trained with the noisy data. As the \textit{Supporting facts} contains cycles, it is difficult for GAT to generalize for a dataset with cycles when it is trained on a dataset without cycles. However, when trained with cycles, GAT learns to attend to all the paths leading to the correct answer. This effect is disastrous when GAT is tested on \textit{Irrelevant facts} which contains dangling paths as GAT still tries to attend to all the paths. Training on \textit{Irrelevant facts} proved to be most beneficial to GAT, as the model now perfectly attends to only \textit{relevant paths}.

Since \textit{Disconnected facts} contains disconnected paths, the message passing function of the graph is unable to forward any information from the disjoint cliques, thereby having superior testing scores throughout several scenarios.

\textbf{Experiments on synthetic placeholders.} In order to further understand the effect of language placeholders on robustness, we performed another set of experiments where we use bAbI (Weston et al., 2015) style simple placeholders (Table 4). We observe a marked increase in performance of all
Figure 6: Systematic Generalizability of different models on CLUTRR-Gen task (having 20% less placeholders and without training and testing placeholder split), when **Left**: trained with \(k = 2\) and \(k = 3\) and **Right**: trained with \(k = 2, 3\) and 4

<table>
<thead>
<tr>
<th>Training</th>
<th>Testing</th>
<th>BiLSTM - Attention</th>
<th>BiLSTM - Mean</th>
<th>RN</th>
<th>MAC</th>
<th>BERT</th>
<th>BERT-LSTM</th>
<th>GAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supporting</td>
<td>Clean</td>
<td>0.38 ± 0.01</td>
<td>0.32 ± 0.01</td>
<td>0.4 ± 0.08</td>
<td>0.45 ± 0.03</td>
<td>0.19 ± 0.06</td>
<td>0.39 ± 0.04</td>
<td>0.92 ± 0.07</td>
</tr>
<tr>
<td></td>
<td>Supporting</td>
<td>0.67 ± 0.06</td>
<td>0.66 ± 0.07</td>
<td>0.68 ± 0.05</td>
<td>0.65 ± 0.02</td>
<td>0.32 ± 0.08</td>
<td>0.57 ± 0.04</td>
<td>0.98 ± 0.03</td>
</tr>
<tr>
<td></td>
<td>Irrelevant</td>
<td>0.44 ± 0.03</td>
<td>0.39 ± 0.02</td>
<td>0.51 ± 0.06</td>
<td>0.46 ± 0.03</td>
<td>0.2 ± 0.01</td>
<td>0.36 ± 0.03</td>
<td>0.5 ± 0.03</td>
</tr>
<tr>
<td></td>
<td>Disconnected</td>
<td>0.31 ± 0.01</td>
<td>0.25 ± 0.06</td>
<td>0.47 ± 0.02</td>
<td>0.41 ± 0.04</td>
<td>0.2 ± 0.02</td>
<td>0.32 ± 0.04</td>
<td>0.92 ± 0.06</td>
</tr>
<tr>
<td>Irrelevant</td>
<td>Clean</td>
<td>0.57 ± 0.05</td>
<td>0.56 ± 0.05</td>
<td>0.46 ± 0.03</td>
<td>0.67 ± 0.05</td>
<td>0.24 ± 0.06</td>
<td>0.46 ± 0.08</td>
<td>0.92 ± 0.08</td>
</tr>
<tr>
<td></td>
<td>Supporting</td>
<td>0.38 ± 0.02</td>
<td>0.31 ± 0.03</td>
<td>0.61 ± 0.07</td>
<td>0.61 ± 0.02</td>
<td>0.27 ± 0.06</td>
<td>0.46 ± 0.08</td>
<td>0.77 ± 0.12</td>
</tr>
<tr>
<td></td>
<td>Irrelevant</td>
<td>0.51 ± 0.06</td>
<td>0.52 ± 0.06</td>
<td>0.5 ± 0.04</td>
<td>0.56 ± 0.02</td>
<td>0.25 ± 0.06</td>
<td>0.53 ± 0.04</td>
<td>0.93 ± 0.03</td>
</tr>
<tr>
<td></td>
<td>Disconnected</td>
<td>0.44 ± 0.02</td>
<td>0.54 ± 0.02</td>
<td>0.55 ± 0.03</td>
<td>0.61 ± 0.02</td>
<td>0.26 ± 0.03</td>
<td>0.45 ± 0.08</td>
<td>0.85 ± 0.08</td>
</tr>
<tr>
<td>Disconnected</td>
<td>Clean</td>
<td>0.45 ± 0.02</td>
<td>0.47 ± 0.02</td>
<td>0.53 ± 0.06</td>
<td>0.5 ± 0.06</td>
<td>0.22 ± 0.06</td>
<td>0.44 ± 0.06</td>
<td>0.75 ± 0.09</td>
</tr>
<tr>
<td></td>
<td>Supporting</td>
<td>0.47 ± 0.03</td>
<td>0.46 ± 0.03</td>
<td>0.54 ± 0.02</td>
<td>0.58 ± 0.02</td>
<td>0.22 ± 0.06</td>
<td>0.38 ± 0.08</td>
<td>0.78 ± 0.12</td>
</tr>
<tr>
<td></td>
<td>Irrelevant</td>
<td>0.47 ± 0.03</td>
<td>0.48 ± 0.02</td>
<td>0.52 ± 0.04</td>
<td>0.51 ± 0.05</td>
<td>0.17 ± 0.04</td>
<td>0.38 ± 0.05</td>
<td>0.56 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>Disconnected</td>
<td>0.57 ± 0.02</td>
<td>0.57 ± 0.02</td>
<td>0.45 ± 0.03</td>
<td>0.4 ± 0.01</td>
<td>0.17 ± 0.04</td>
<td>0.47 ± 0.04</td>
<td>0.96 ± 0.03</td>
</tr>
<tr>
<td>Average</td>
<td>Clean</td>
<td>0.47 ± 0.02</td>
<td>0.46 ± 0.02</td>
<td>0.53 ± 0.07</td>
<td>0.51 ± 0.07</td>
<td>0.23 ± 0.07</td>
<td>0.43 ± 0.05</td>
<td>0.82 ± 0.08</td>
</tr>
</tbody>
</table>

Table 3: Testing the robustness of the various models when trained various types of noisy facts and evaluated on other noisy / clean facts. The types of noise facts (supporting, irrelevant and disconnected) are defined in Section 3.5 of the main paper.

<table>
<thead>
<tr>
<th>Training</th>
<th>Testing</th>
<th>BiLSTM - Attention</th>
<th>BiLSTM - Mean</th>
<th>RN</th>
<th>MAC</th>
<th>BERT</th>
<th>BERT-LSTM</th>
<th>GAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supporting</td>
<td>Clean</td>
<td>0.96 ± 0.06</td>
<td>0.97 ± 0.01</td>
<td>0.88 ± 0.05</td>
<td>0.94 ± 0.02</td>
<td>0.48 ± 0.04</td>
<td>0.57 ± 0.04</td>
<td>0.92 ± 0.07</td>
</tr>
<tr>
<td></td>
<td>Supporting</td>
<td>0.96 ± 0.06</td>
<td>0.96 ± 0.07</td>
<td>0.97 ± 0.01</td>
<td>0.97 ± 0.02</td>
<td>0.75 ± 0.07</td>
<td>0.88 ± 0.05</td>
<td>0.98 ± 0.04</td>
</tr>
<tr>
<td></td>
<td>Irrelevant</td>
<td>0.92 ± 0.06</td>
<td>0.93 ± 0.06</td>
<td>0.91 ± 0.03</td>
<td>0.91 ± 0.01</td>
<td>0.56 ± 0.04</td>
<td>0.54 ± 0.01</td>
<td>0.92 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>Disconnected</td>
<td>0.8 ± 0.04</td>
<td>0.83 ± 0.04</td>
<td>0.76 ± 0.02</td>
<td>0.86 ± 0.01</td>
<td>0.27 ± 0.06</td>
<td>0.42 ± 0.01</td>
<td>0.92 ± 0.06</td>
</tr>
<tr>
<td>Irrelevant</td>
<td>Clean</td>
<td>0.62 ± 0.03</td>
<td>0.64 ± 0.04</td>
<td>0.85 ± 0.08</td>
<td>0.88 ± 0.07</td>
<td>0.53 ± 0.06</td>
<td>0.44 ± 0.06</td>
<td>0.92 ± 0.04</td>
</tr>
<tr>
<td></td>
<td>Supporting</td>
<td>0.66 ± 0.04</td>
<td>0.64 ± 0.05</td>
<td>0.69 ± 0.06</td>
<td>0.76 ± 0.04</td>
<td>0.42 ± 0.08</td>
<td>0.43 ± 0.08</td>
<td>0.77 ± 0.12</td>
</tr>
<tr>
<td></td>
<td>Irrelevant</td>
<td>0.89 ± 0.14</td>
<td>0.86 ± 0.11</td>
<td>0.74 ± 0.11</td>
<td>0.78 ± 0.09</td>
<td>0.61 ± 0.1</td>
<td>0.83 ± 0.08</td>
<td>0.93 ± 0.03</td>
</tr>
<tr>
<td></td>
<td>Disconnected</td>
<td>0.64 ± 0.02</td>
<td>0.62 ± 0.03</td>
<td>0.72 ± 0.08</td>
<td>0.73 ± 0.04</td>
<td>0.41 ± 0.04</td>
<td>0.61 ± 0.03</td>
<td>0.85 ± 0.06</td>
</tr>
<tr>
<td>Disconnected</td>
<td>Clean</td>
<td>0.9 ± 0.01</td>
<td>0.82 ± 0.02</td>
<td>0.94 ± 0.02</td>
<td>0.93 ± 0.03</td>
<td>0.6 ± 0.07</td>
<td>0.64 ± 0.02</td>
<td>0.75 ± 0.07</td>
</tr>
<tr>
<td></td>
<td>Supporting</td>
<td>0.87 ± 0.04</td>
<td>0.85 ± 0.05</td>
<td>0.85 ± 0.08</td>
<td>0.88 ± 0.04</td>
<td>0.54 ± 0.08</td>
<td>0.5 ± 0.01</td>
<td>0.78 ± 0.12</td>
</tr>
<tr>
<td></td>
<td>Irrelevant</td>
<td>0.87 ± 0.04</td>
<td>0.85 ± 0.03</td>
<td>0.83 ± 0.03</td>
<td>0.87 ± 0.02</td>
<td>0.59 ± 0.09</td>
<td>0.58 ± 0.09</td>
<td>0.56 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>Disconnected</td>
<td>0.91 ± 0.01</td>
<td>0.91 ± 0.02</td>
<td>0.8 ± 0.07</td>
<td>0.71 ± 0.11</td>
<td>0.49 ± 0.04</td>
<td>0.79 ± 0.11</td>
<td>0.96 ± 0.03</td>
</tr>
<tr>
<td>Average</td>
<td>Clean</td>
<td>0.83 ± 0.01</td>
<td>0.82 ± 0.01</td>
<td>0.83 ± 0.07</td>
<td>0.84 ± 0.06</td>
<td>0.58 ± 0.07</td>
<td>0.60 ± 0.01</td>
<td>0.82 ± 0.01</td>
</tr>
</tbody>
</table>

Table 4: Testing the robustness to toy placeholders of the various models when trained various types of noisy facts and evaluated on other noisy / clean facts. The types of noise facts (supporting, irrelevant and disconnected) are defined in Section 3.5 of the main paper.

NLU models, where they significantly decrease the gap between their performance with that of GAT, even outperforming GAT on various settings. This shows the significance of using paraphrased placeholders in devising the complexity of the dataset.

### 1.5 Comparison among different entity embedding policies

In Cloze style reading comprehension tasks, it is sometimes customary to choose UNK embeddings
for entity placeholders. (Chen et al., 2016) In our task, however, choosing UNK embeddings for entities is not feasible as the query involves two entities themselves. During preprocessing of our dataset, we convert the entity names into a Cloze-style setup where each entity is replaced by @entity-n token. However, one has to be careful not to assign tokens in the same order for all the stories, which will lead to obvious overfitting since the models will learn to work around positional markers as shown in Chen et al. (2016). Therefore, we randomize the Cloze-style entities themselves for each story. We experimented with three different policies of choosing the entity embeddings:

1. **Fixed Random Embeddings**: One simple and intuitive choice is to assign a random embedding to each entity and keep it fixed throughout the training. During our data-processing pipeline, we ensure that all the entity tokens are randomized using a pool of entity tokens, hence the chances of a model learning to exploit the positional markers are slim.

2. **Randomized Random Embeddings**: We can go one step further and randomize the random embeddings at each epoch. This aggressive strategy does not let the model learn any positional markings at all, however it might hamper the learning ability of models as the entity representations are changing arbitrarily.

3. **Learned Random Embeddings**: Since our data pre-processing pipeline randomly assigns the entities on each story, we can as well learn a pool of n entities, from which a subset is always used to replace the entities.

We chose to report all experiments with respect to fixed random embeddings. We compared different embedding policies with respect to the Systematic Generalization task. We show a comparison between the Bidirectional LSTM and GAT in Figure 7. We see that the fixed embedding policy has better Systematic Generalization score, although the advantage is minor compared to the other schemes. For GAT, the advantage is practically nil for the different schemes which shows that a Graph Neural Network performs inductive reasoning in the same manner irrespective of the initial node embedding representation.

### 1.6 AMT Data collection process

We use ParlAI (Miller et al., 2017) Mturk interface to collect paraphrases from the users. Specifically, given a set of facts, we ask the users to paraphrase the facts into a story. The users (turkers) are free to construct any story they like as long as they mention all the entities and all the relations among them. We also provide the head $H$ of the clause as an inferred relation and specifically instruct the users to not mention it in the paraphrased story. In order to evaluate the paraphrased stories, we ask the turkers to peer review a story paraphrased by a different turker. Since there are two tasks - paraphrasing a story and rating a story - we choose to pay 0.5$ for each annotation. A sample task description in our MTurk interface is as follows:

**AMT Data collection process**

In this task, you will need to write a short, simple story based on a few facts. **It is crucial that the story mentions each of the given facts at least once.** The story does not need to be complicated! It just needs to be grammatical and mention the required facts.

After writing the story, you will be asked to evaluate the quality of a generated story (based on a different set of facts). **It is crucial that you check whether the generated story mentions each of the required facts.**

**Example of good and bad stories: Good Example**

**Facts to Mention**
- John is the father of Sylvia.
- Sylvia has a brother Patrick.

**Implied Fact**: John is the father of Patrick.

**Written story**
John is the proud father of the lovely Sylvia. Sylvia has a love-hate relationship with her brother Patrick.

**Bad Example**

**Facts to Mention**
- Vincent is the son of Tim.
- Martha is the wife of Tim.

**Implied Fact**: Martha is Vincent’s mother.

**Written story**
Vincent is married at Tim and his mother is Martha.

**The reason the above story is bad:**
- This story is bad because it is nonsense / ungrammatical.
- This story is bad because it does not mention the proper facts.
- This story is bad because it reveals the implied fact.
Table 5: Human performance accuracies on CLUTRR dataset. Humans are provided the Clean-Generalization version of the dataset, and we test on two scenarios: when a human is given limited time to solve the task, and when a human is given unlimited time to solve the task. Regardless of time, our evaluators provide a score of difficulty of individual puzzles.

To ensure that the turkers are providing high-quality annotations without revealing the inferred fact, we also launch another task to ask the turkers to rate three annotations to be either good or bad which are provided by a set of different turkers. We pay 0.2$ for each HIT consisting of three reviews. This helped to remove logical and grammatical inconsistencies to a large extent. Based on the reviews, 79% of the collected paraphrases passed the peer-review sanity check where all the reviewers agree on the quality. This subset of the placeholders is used in the benchmark. A sample of programmatically generated dataset for clause length of \( k = 2 \) to \( k = 6 \) is provided in the tables 7 to 11.

1.7 Human Evaluation

We performed a human evaluation study to analyze the difficulty of our proposed benchmark suite, which is provided in Table 5. We perform the evaluation in two scenarios: first a time-limited scenario where we ask AMT Turkers to solve the puzzle in a fixed time. Turkers were provided a maximum time of 30 mins, but they solved the puzzles in an average of 1 minute 23 seconds. Secondly, we use another set of expert evaluators who are given ample time to solve the tasks. Not surprisingly, if a human being is given ample time (experts took an average of 6 minutes per puzzle) and a pen and a paper to aid in the reasoning, they get all the relations correct. However, if an evaluator is short of time, they might miss important details on the relations and perform poorly. Thus, our tasks require active attention.

In both cases, we asked Turkers and our expert human evaluators to rate the difficulty of a given task in a Likert scale of 1-5, where 1 corresponds to very easy and 5 corresponds to very hard perceived difficulty. This score increases as we increase the complexity of the task by increasing the relations, thereby suggesting that a human being perceives similar difficulty while solving for larger relation tasks. However, since a human being is a systematic learner, given enough time they can solve all puzzles with perfect accuracy. We set aside the task of testing noisy scenarios of CLUTRR to human evaluators as future work.

2 Supplemental Material

To promote reproducibility, we follow the guidelines proposed by the Machine Learning Reproducibility Checklist and release the following information regarding the experiments conducted by our benchmark suite.

2.1 Details of datasets used

A downloadable link to the datasets used can be found here. Details of the individual datasets can

---

7 Machine Learning Reproducibility Checklist
8 Dataset link in Google Drive
be found in Table 6. For all experiments, we use 10,000 training examples and a 100 testing example for each testing scenario. We split the training data 80-20 into a dev set randomly on each run.

2.2 Details of Hyperparameters used

For all models, the common hyperparameters used are: Embedding dimension: 100 (except BERT based models), Optimizer: Adam, Learning rate: 0.001, Number of epochs: 100, Number of runs: 10. Specific model-based hyperparameters are given as follows:

- **Bidirectional LSTM**: LSTM hidden dimension: 100, # layers: 2, Classifier MLP hidden dimension: 200

- **Relation Networks**: $f_{\theta_1}$: 256, $f_{\theta_2}$: 64, $g_{\theta}$: 64

- **MAC**: # Iterations: 6, shareQuestion: True, Dropout - Memory, Read and Write: 0.2

- **Relational Recurrent Networks**: Memory slots: 2, Head size: 192, Number of heads: 4, Number of blocks: 1, forget bias: 1, input bias: 0, gate style: unit, key size: 64, # Attention layers: 3, Dropout: 0

- **BERT**: Layers: 12, Fixed pretrained embeddings from bert-base-uncased using Pytorch HuggingFace BERT repository, Word dimension: 768, appended with a two-layer MLP for final prediction.

- **BERT-LSTM**: Same parameters as above, with a two-layer unidirectional LSTM encoder on top of BERT word embeddings.

- **GAT**: Node dimension: 100, Message dimension: 100, Edge dimension: 20, number of rounds: 3
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Variant - Training</th>
<th>Variant - Testing</th>
<th>Training Clause length</th>
<th>Testing Clause length</th>
</tr>
</thead>
<tbody>
<tr>
<td>data_08907f88</td>
<td>Clean - Generalization</td>
<td>Clean - Generalization</td>
<td>(k = 2, 3)</td>
<td>(k = 2, 3, ..., 10)</td>
</tr>
<tr>
<td>db9b8f04</td>
<td>Clean, Supporting, Irrelevant, Disconnected</td>
<td>Clean, Supporting, Irrelevant, Disconnected</td>
<td>(k = 2, 3)</td>
<td>(k = 2, 3)</td>
</tr>
<tr>
<td>data_7c5b0e70</td>
<td>Supporting</td>
<td>Clean, Supporting, Irrelevant, Disconnected</td>
<td>(k = 2, 3)</td>
<td>(k = 2, 3)</td>
</tr>
<tr>
<td>data_523348e6</td>
<td>Irrelevant</td>
<td>Clean, Supporting, Irrelevant, Disconnected</td>
<td>(k = 2, 3)</td>
<td>(k = 2, 3)</td>
</tr>
<tr>
<td>data_523ec3e</td>
<td>Disconnected</td>
<td>Clean, Supporting, Irrelevant, Disconnected</td>
<td>(k = 2, 3)</td>
<td>(k = 2, 3)</td>
</tr>
</tbody>
</table>

Table 6: Details of publicly released data

Figure 8: Amazon Mechanical Turk Interface built using ParlAI which was used to collect data as well as peer reviews.
### Table 7: Snapshot of puzzles in the dataset for k=2

<table>
<thead>
<tr>
<th>Puzzle</th>
<th>Question</th>
<th>Gender</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Charles’s son Christopher entered rehab for the ninth time at the age of thirty. Randolph had a nephew called Christopher who had n’t seen for a number of years.</td>
<td>Randolph is the ___ of Charles</td>
<td>Charles:male, Christopher:male, Randolph:male</td>
<td>brother</td>
</tr>
<tr>
<td>Randolph and his sister Sharon went to the park. Arthur went to the baseball game with his son Randolph</td>
<td>Sharon is the ___ of Arthur</td>
<td>Arthur:male, Randolph:male, Sharon:female</td>
<td>daughter</td>
</tr>
<tr>
<td>Frank went to the park with his father, Brett. Frank called his brother Boyd on the phone. He wanted to go out for some beers.</td>
<td>Brett is the ____ of Boyd</td>
<td>Boyd:male, Frank:male, Brett:male</td>
<td>father</td>
</tr>
</tbody>
</table>

### Table 8: Snapshot of puzzles in the dataset for k=3

<table>
<thead>
<tr>
<th>Puzzle</th>
<th>Question</th>
<th>Gender</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roger was playing baseball with his sons Sam and Leon. Sam had to take a break though because he needed to call his sister Robin.</td>
<td>Leon is the ___ of Robin</td>
<td>Robin:female, Sam:male, Roger:male, Leon:male</td>
<td>brother</td>
</tr>
<tr>
<td>Elvira and her daughter Nancy went shopping together last Monday and they bought new shoes for Elvira’s kids. Pedro and his sister Allison went to the fair. Pedro’s mother, Nancy, was out with friends for the day.</td>
<td>Elvira is the ____ of Allison</td>
<td>Allison:female, Pedro:male, Nancy:female, Elvira:female</td>
<td>grandmother</td>
</tr>
<tr>
<td>Roger met up with his sister Nancy and her daughter Cynthia at the mall to go shopping together. Cynthia’s brother Pedro was going to be the star in the new show.</td>
<td>Pedro is the ____ of Roger</td>
<td>Roger:male, Nancy:female, Cynthia:female, Pedro:male</td>
<td>nephew</td>
</tr>
</tbody>
</table>
Table 9: Snapshot of puzzles in the dataset for \( k=4 \)

<table>
<thead>
<tr>
<th>Puzzle</th>
<th>Question</th>
<th>Gender</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Celina has been visiting her sister, Fran all week. Fran is also the daughter of Bethany. Ronald loves visiting his aunt Bethany over the weekends. Samuel’s son Ronald entered rehab for the ninth time at the age of thirty.</td>
<td>Celina is the ____ of Samuel</td>
<td>Samuel: male, Ronald: male, Bethany: female, niece Fran: female, Celina: female</td>
<td></td>
</tr>
<tr>
<td>Celina adores her daughter Bethany. Bethany loves her very much, too. Jackie called her mother Bethany to let her know she will be back home soon. Thomas was helping his daughter Fran with her homework at home. Afterwards, Fran and her sister Jackie played Xbox together.</td>
<td>Celina is the ____ of Thomas</td>
<td>Thomas: male, Fran: female, Jackie: female, daughter Bethany: female, Celina: female</td>
<td></td>
</tr>
<tr>
<td>Raquel is Samuel’s daughter and they go shopping at least twice a week together. Kenneth and her mom, Theresa, had a big fight. Theresa’s son, Ronald, refused to get involved. Ronald was having an argument with her sister, Raquel.</td>
<td>Samuel is the ____ of Kenneth</td>
<td>Kenneth: male, Theresa: female, Ronald: male, father Raquel: female, Samuel: male</td>
<td></td>
</tr>
</tbody>
</table>
### Table 10: Snapshot of puzzles in the dataset for k=5

<table>
<thead>
<tr>
<th>Puzzle</th>
<th>Question</th>
<th>Gender</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steven’s son is Bradford. Bradford and his father always go fishing together on Sundays and have a great time together. Diane is taking her brother Brad out for a late dinner. Kristin, Brad’s mother, is home with a cold. Diane’s father Elmer, and his brother Steven, all got into the rental car to start the long cross-country roadtrip they had been planning.</td>
<td>Bradford is the _____ of Kristin</td>
<td>Kristin:female, Brad:male, Elmer:male, Steven:male, Bradford:male</td>
<td>nephew</td>
</tr>
<tr>
<td>Elmer went on a roadtrip with his youngest child, Brad. Lena and her sister Diane are going to a restaurant for lunch. Lena’s brother Brad is going to meet them there with his father Elmer Brad can’t stand his unfriendly aunt Lizzie.</td>
<td>Lizzie is the _____ of Diane</td>
<td>Diane:female, Lena:female, Elmer:male, Lizzie:female</td>
<td>aunt</td>
</tr>
<tr>
<td>Ira took his niece April fishing Saturday. They caught a couple small fish. Ronald was enjoying spending time with his parents, Damion and Claudine. Damion’s other son, Dennis, wanted to come visit too. Dennis often goes out for lunch with his sister, April.</td>
<td>Ira is the _____ of Claudine</td>
<td>Claudine:female, Ronald:male, Damion:male, Dennis:male, April:female, Ira:male</td>
<td>brother</td>
</tr>
</tbody>
</table>
Table 11: Snapshot of puzzles in the dataset for k=6

<table>
<thead>
<tr>
<th>Puzzle</th>
<th>Question</th>
<th>Gender</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mario wanted to get a good gift for his sister, Marianne. Jean and her sister Darlene were going to a party held by Jean’s mom, Marianne. Darlene invited her brother Roy to come, too, but he was too busy. Teri and her father, Mario, had an argument over the weekend. However, they made up by Monday. Agnes wants to make a special meal for her daughter Teri’s birthday.</td>
<td>Roy is the _____ of Agnes</td>
<td>Agnes:female, Teri:female, Mario:male, Marianne:female, Jean:female, Darlene:female, Roy:male</td>
<td></td>
</tr>
<tr>
<td>Robert’s aunt, Marianne, asked Robert to mow the lawn for her. Robert said he couldn’t because he had a bad back. William’s parents, Brian and Marianne, threw him a surprise party for his birthday. Brian’s daughter Jean made a mental note to be out of town for her birthday! Agnes’s biggest accomplishment is raising her son Robert. Jean is looking for a good gift for her sister Darlene.</td>
<td>Darlene is the _____ of Agnes</td>
<td>Agnes:female, Robert:male, Marianne:female, William:male, Brian:male, Jean:female, Darlene:female</td>
<td></td>
</tr>
<tr>
<td>Sharon and her brother Mario went shopping. Teri, Mario’s daughter, came too. Agnes, Annie’s mother, is unhappy with Robert. She feels her son is cruel to Annie’s sister Teri, and she wants Robert to be nicer. Robert’s sister, Nicole, participated in the dance contest.</td>
<td>Nicole is the _____ of Sharon</td>
<td>Sharon:female, Mario:male, Teri:female, Annie:female, Agnes:female, Robert:male, Nicole:female</td>
<td></td>
</tr>
</tbody>
</table>