# Dynamic Neural-Symbolic Reasoner on Commonsense Knowledge

*Social commonsense reasoning requires capitalisation on Commonsense Knowledge Graphs (CKG) and exploitation of explicit and implicit relationships among events, in order to draw conclusion. However, some characteristics of such knowledge bases poses new challenges. The non-canonicalised free-form text representation of events in CKGs has resulted in large-scale sparse graphs. In addition, a huge and fast-growing number of social situations require models to be capable of reasoning over diverse and unseen situations. In this paper, we present a unified dynamic neural-symbolic reasoner to address the tasks of CKG completion and zero-shot Commonsense Question Answering (CQA). During training, our model learns transition probabilities of logical rules for multi-hop reasoning over CKGs. In addition to providing interpretable explanations, the learnt logical rule transition patterns help to generalise prediction to address the task of CQA. The empirical results show that our model significantly outperforms state-of-the-art models in both tasks.*

## 1. Introduction

Commonsense reasoning is the ability to make presumptions concerning an ordinary context. This process encourages anticipation and reasoning over a commonly known background knowledge to make a plausible conclusion (Davis and Marcus 2015). While this process helps humans manage day-to-day encounters trivially, empowering current Artificial Intelligent models with this ability yet to be addressed (Sap et al. 2020).

Recent fast-growing interests in endowing AI systems with such human-like capability have focused on using background Commonsense Knowledge Graphs (CKGs) to develop commonsense reasoning engines (Malaviya et al. 2020; Moghimifar et al. 2021a). The performance of such engines is often evaluated by their ability in inferring a follow-up event given a narrative and a specific dimension. Such reasoning process requires models to identify underlying implicit relations and account for a chain of events to draw a plausible conclusion.

However, some characteristics of commonly used CKGs, such as ATOMIC and ConceptNet (Sap et al. 2019a; Speer, Chin, and Havasi 2017), pose challenges in performing multi-hop reasoning process. Firstly, the facts in CKGs are encoded in form of arbitrary phrases. Therefore conceptually related nodes are represented in various forms which results in having large-scale sparse knowledge graphs. For instance, while the nodes “Alex thanks Jesse" and “Alex is grateful towards Jesse" carry the same semantic meaning, are represented in various formats. This non-canonicalised representation of nodes in CKGs has resulted in having *large-scale sparse* knowledge bases (Table 1). Secondly, the dynamic world of commonsense knowledge introduces new facts to CKGs frequently, which challenges the reasoning process on unseen events. This property encourages commonsense reasoning engines to be able to generalise the inference process to previously unseen context. Consequently, current state-of-the-art models in this area fail to perform adequately when there is a distributional shift between the given narrative and nodes in the CKG.

Recent attempts in addressing the task of Commonsense Question Answering (CQA) have also focused on leveraging background CKGs as a source of information (Bosselut

| Dataset | #Nodes | #Edges | Avg. In-degree | Density | Unseen Nodes | Unseen Edges | #Relations |
|---|---|---|---|---|---|---|---|
| ATOMIC | 382823 | 785952 | 2.25 | 1.6e-5 | 38.36% | 27.91% | 9 |
| ConceptNet-100k | 80994 | 102400 | 1.25 | 9.0e-6 | 11% | 8% | 34 |

Table 1: Statistics on ATOMIC and ConceptNet-100k. Unseen Nodes is the ratio of the nodes in test set that are not in train set to all of the nodes in test set. Unseen edges is the ratio of edges where either the head or tail nodes are not in train set to the number of all edges in test set.

and Choi 2019; Moghimifar et al. 2020). For this purpose, the given commonsense context is mapped to one the nodes of the CKG, and the reasoning process is delivered on the CKG, until a most plausible answer is identified. However, the reasoning process of the current state-of-the-art approaches in this area is limited by exhausting all possible connections between nodes, resulting in higher computational complexity and less robustness of the performance (Bosselut and Choi 2019; Moghimifar et al. 2020).

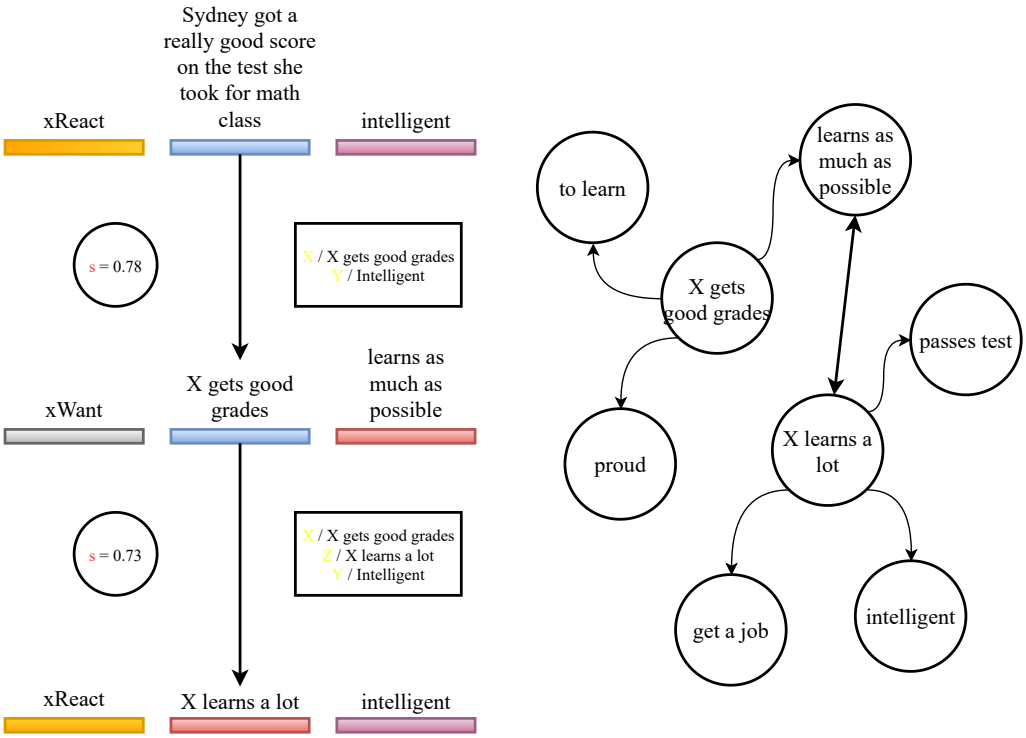


Figure 1: A visual representation of a sample query from SocialIQA test set. The context of this data is *Sydney got a really good score on the test she took for math class.* and Sydney's feeling afterwards has been asked, with the correct answer of *Intelligent*. Initially, the reasoner starts with the query: *question(context,candidate answer)*. Using the learnt rule of `xReact(X,Z) :- xWant(X,Z), xReact(Z,Y)` the query is updated at each step, until the query is proven. The box above each arrow represent unification of variables with nodes of CKG, and the box below each arrow represent unification score. The graph on the right side represents a subgraph of ATOMIC neighbouring nodes of the queries.

To address the above-mentioned limitations, we propose a unified neural-symbolic commonsense reasoning model based on forward-chaining. Inspired by traditional theorem prover, the proposed model replaces discrete symbols with continuous embedding representation and by leveraging weak unification on a forward-chaining approach, performs multi-hop reasoning. For this purpose, during training, our model learns a set of logical rules over a CKG, and uses the learned rules on inference time to generalise to new unseen presented context. To this end, our proposed model dynamically uses an additional source of knowledge, a generative SEQ2SEQ model, when there is a distributional shift between the context and the CKG. Figure 1 illustrates a high-level process of our proposed model. We evaluate the performance of our model on the task of CKG completion and zero-shot Commonsense Question Answering (CQA). The experimental results suggest that our model outperforms current state-of-the-art models.

## 2. Related Works

### 2.1 Commonsense Knowledge Graph

Knowledge Graphs (KG) have been used significantly to provide external knowledge for machines on the comprehension task. The motivation is to incorporate human knowledge to help artificial intelligence solve complex tasks. KG is a structural knowledge representation containing human facts, entities, relations and semantic descriptions. The history of KG can be back to the time of knowledge base (KB), where orginal KBs were only made up of pre-defined rules and without structural information. KBs later were developed into the form of combining human knowledge with ontologies, such as WordNet (Fellbaum 2010), DBpedia (Auer et al. 2007) and YAGO (Suchanek, Kasneci, and Weikum 2007). However, these KBs were still limited due to the shallow knowledge and failed to aid AI systems reason over complex task. Therefore, the community have made huge effort (Paulheim 2017; Ehrlinger and Wöß 2016; Wang et al. 2017) on the formal definition of the essential structure of KGs. According to (Ji et al. 2021), current KGs need to integrate information to an ontology, be able to facilitate a reasoner to derive new knowledge via multiple relations.

Recent growing interests in commonsense reasoning points to the direction of constructing large commonsense knowledge graphs (CKG) containing various daily knowledge. Conventionally, the edges and nodes in these KGs represent relations and content, respectively. There are mainly two types of CKGs lying in the commonsense reasoning domain, concept-centric and event-centric, where concept-centric CKGs focus on relations among daily entities (e.g. fruits and animals), and event-centric CKGs are more applicable to daily human-involved events. Several concept-centric CKGs (Speer, Chin, and Havasi 2017; Carlson et al. 2010; Wu et al. 2012) have shown the effectiveness on language comprehension tasks. While those CKGs are more centered around taxonomic knowledge, event-centric CKGs consist of inferential knowledge. Some effort in creating event-centric CKGs (Sap et al. 2019a; Zhang et al. 2020; Mostafazadeh et al. 2020) have been exploited to generate more meaningful sentences. Later, Hwang et al. (2020) proposed a unified CKGs combining both concept-centric and event-centric information together, based on the previous if-else event-centric ATOMIC (Sap et al. 2019a).

### 2.2 Commonsense Knowledge Graph Completion

KG completion task is designed for evaluating the performance of KG embedding, which is generally applied to the daily scenarios, such as question answering, search

and recommendation. In these areas, normal discrete KGs are not applicable due to the unscalable knowledge. KG embedding is one of solution to extend the discrete knowledge to the continuous space. Most existing KG embedding methods focus on open-domain factoid knowledge graphs. However, few of them also target on CKGs, such as FB15k (Toutanova and Chen 2015), ConceptNet (Speer, Chin, and Havasi 2017) and ATOMIC (Sap et al. 2019a).

Previous works on knowledge base completion task, mostly focused on learning node to address this task (Yang et al. 2014; Dettmers et al. 2018). Entities and relations are embedded in a complex space, and the plausibility of a triple is calculated using a scoring function in these techniques. More advanced approaches embedded the node and graph representation into complex space (Sun et al. 2018; Shang et al. 2019; Trouillon et al. 2016). However, these methods perform poorly on CKGs, due to the sparsity of these graphs. Malaviya et al. (2020) leveraged graph network embeddings and language models, by considering the structural and contextual properties of CKGs, to estimate a node. However, this method depends on observing all the nodes in the training session. (Wang et al. 2021) defined the inductive learning problem on CKG completion, and hence created a framework called InductivE, greatly outperforming previous works on the same task. Moghimifar et al. (2021b) proposed a method based on backward chaining to generalise the inference to unseen nodes. Despite that the logic rules in their framework help to provide interpretable explanation and outperform previous CKG completion methods, they are still quite shallow, which fail to capture relation between predicates.

### 2.3 Zero-shot Commonsense Question Answering

The recent surge in addressing the task of commonsense question answering (CQA) has resulted in using pretrained language models to capture the association between context and the correct answer (Sap et al. 2019b; Zellers et al. 2018). However, the dynamic nature of commonsense world motivated proposing unsupervised models (Bosselut and Choi 2019), with using background knowledge. Similar ideas on utilizing external knowledge also improve greatly on zero-shot CQA. Shwartz et al. (2020) proposed a prompt-based Self-Talk procedure to answer questions via language models training on CKGs, without any parallel CQA training data. Self-Talk model specifically use COMET (Bosselut et al. 2019) to generate commonsense background knowledge for selecting the best answer based on the given questions and context. Other approaches (Banerjee and Baral 2020; Moghimifar et al. 2020) have been explored to generalise unseen questions on CQA tasks. Zero-shot CQA generally tests the model generalisation on some complex context comprehension tasks via learning some related commonsense knowledge. Nevertheless, these approaches exhaust all possible paths to finds the answer, which results in poor performance.

## 3. Approach

Given a CKG in form of $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ is the set of nodes, and $\mathcal{E}$ is the set of edges of the graph $\mathcal{G}$. $\mathcal{G}$ consists of facts in form of $r(h, t)$ where, $h, t \in \mathcal{N}$ are called head and tail of a relation, and $r \in \mathcal{E}$ represents the relation between head and tail. The goal of the reasoning engine for a given query $r_q(h_q, ?)$ is to find the most plausible target entity $t \in \mathcal{E}$, by finding the most plausible reasoning path $\boldsymbol{Z}$, which starts from $h_q$, along the

target relation until it reaches $t$. Therefore, the object of the reasoner is to find the most plausible answer by solving the following optimisation problem:

$$\arg \max_{\boldsymbol{t} \in \mathcal{H}, \boldsymbol{Z} \in \mathcal{Z}} \log Pr(\boldsymbol{t}|\boldsymbol{Z}) Pr(\boldsymbol{Z}|\boldsymbol{h_q}, r_q, \mathcal{G}) \tag{1}$$

where $\mathcal{Z}$ is the set of possible path on $\mathcal{G}$ starting from $h_q$.

The local distribution of $Pr(\boldsymbol{Z}|\boldsymbol{h_q}, r_q, \mathcal{G})$ can be represented as:

$$Pr(\boldsymbol{Z}|h_q, r_q, \mathcal{G}) = min\{Pr(h_q|r_q, \mathcal{G}), Pr(\boldsymbol{z}_t|\boldsymbol{z}_1, r_1, \mathcal{G}), \ldots, Pr(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, r_{t-1}, \mathcal{G})\} \tag{2}$$

where $\boldsymbol{z_{t-1}}$ and $\boldsymbol{r_{t-1}}$ are the node and relation in previous time step, respectively. The term $Pr(h_q|r_q, \mathcal{G})$ represents the mapping process of a given context to one of the nodes in the CKG, and $Pr(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, r_{t-1}, \mathcal{G})$ represents the probably associated with each reasoning step taken by the model.

For estimating equation 1 a rule $\mathcal{R}$, in form of $r_q(X, Z) :- r_0(X, Y_0), ..., r_k(Y_{k-1}, Z)$ is applied to next step, where capitalised letters denote variables, $r_q(X, Z)$ is the rule head, and the rule body is a conjunction of atoms. By unifying atoms in $\mathcal{G}$ using the rule $\mathcal{R}$, a reasoning path such as $r_q(h_q, t_k) :- r_0(h_0, t_0), r_1(h_1, t_1), ..., r_k(h_k, t_k)$, is obtained where $r_q(h_q, t_k)$ can be inferred.

At each time step, the rightmost atom of the rule is used as query $r_{k-1}(h_{k-1}, X)$. The representation of the head of the query, $h_{k-1}$ is then used to retrieve the potential unified candidates from $\mathcal{G}$. Since the nodes in $\mathcal{G}$ are represented as a sequence of words, we encode nodes with a pre-trained language model (Devlin et al. 2019) into embeddings. This representation is achieved by converting the node into $[CLS] + node + [SEP]$ and feeding it to the language model, and getting the representation of $[CLS]$ from the last layer of the model. We retrieve the $k$ nearest neighbour of $h_{k-l}$ by indexing all nodes in the knowledge graph using FAISS (Johnson, Douze, and Jégou 2019), and we collect all the triples from $\mathcal{G}$ where $h_{k-l}$ is in the head, forming a subset of $\mathcal{G}$, named $\mathcal{C}$.

Inspired by (Sessa 2002) and (Moghimifar et al. 2021a), to enforce a continuous relaxation of node representation, we adopt a weak unification approach to find $X$ in query $r_{k-1}(h_{k-1}, X)$. To this end, we replace $X$ with all possible tail nodes from $\mathcal{C}$, named hypothesis $\mathcal{H}$. We measure the similarity between $\mathcal{C}$ and $\mathcal{H}$, forming a matrix $\mathbf{U} \in \mathbb{R}^{|\mathcal{C}| \times |\mathcal{H}|}$. The final $k$ candidates for unification is then calculated by $\max_j \mathbf{U}_{ij}$.

Upon identifying the top-$k$ candidates for replacing $X$, the rule $\mathcal{R}$ is then updated by appending $r_t(c_k, X)$, where $c_k$ is the top-$k$ candidates. Proposed by Moghimifar et al. (Moghimifar et al. 2021a) the relation $r_k$ is estimated by the following relation prediction module:

$$P_{\boldsymbol{\theta_f}}(r_k|r_{k-1}, k) = \sigma(\boldsymbol{f_\theta}([r_{k-1}; k]).\mathbf{W} + b) \tag{3}$$

where $\boldsymbol{\theta_f} := \{\mathbf{W}, b\}$ contains the module's parameters, and $\sigma$ is the sigmoid function. The relation predictor aims to generalise relation co-occurrence patterns in rules. It is implemented by using a neural networks with two blocks of hidden layers, followed by a softmax layer. Each block is composed of a linear layer and a ReLU layer.

However, using Equation 3 fails to capture the graphical structure neighbouring the query node at each step. To overcome this problem, we propose a model based Graph Convolutional Networks (GCN) where the relation $r_k$ is estimated by:

$$P_{\boldsymbol{\theta_f}}(r_k|\mathcal{C}, r_{k-1}, t_{k-1}) = \sigma(\boldsymbol{f_\theta}([\boldsymbol{f_{\theta_c}}; r_{k-1}; t_{k-1}]).\mathbf{W} + b) \tag{4}$$

where $\boldsymbol{\theta_f} := \{\mathbf{W}, b\}$ contains the Rule creation module's parameters, and $\sigma$ is the sigmoid function. $\boldsymbol{f_{\theta_c}}$ is the representation of the graph $\mathcal{C}$ neighbouring node $t_{k-1}$. The graph representation is retrieved by:

$$\boldsymbol{f_{\theta_c}} = FFNN(f^{pool}(GCN(\mathcal{C}))) \tag{5}$$

where $f^{pool} : \mathbb{R}^{dxn} \rightarrow \mathbb{R}^d$ is a pooling function, which generates representation for nodes of the graph $\mathcal{C}$, and FFNN is the feed forward neural network. The GCN is the graph convolutional network function over the adjacency matrix $\mathbb{A}$ of graph $\mathcal{G}$, by which the node representation at layer $l$ can be estimated by:

$$\boldsymbol{h}_i^{(l)} = \boldsymbol{f}^{\text{actv}}(\sum_{j=1}^{n} \tilde{A}_{ij}\mathbf{W}^{(l)}\boldsymbol{h}_j^{(l-1)}/d_i + \boldsymbol{b}^{(l)}) \tag{6}$$

where $\tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}$, with $\boldsymbol{I}$ the identity matrix, $\boldsymbol{f}^{\text{actv}}$ an activation function (i.e., element-wise RELU), $\boldsymbol{b}^{(l)}$ the bias vector, $\mathbf{W}^{(l)}$ the weight matrix, and $d_i = \sum_{j=1}^{n} \tilde{A}_{ij}$ the degree of node $i$.

The process of weak unification and appending atoms to the rule on the fly is continued until a predefined maximum depth is reached. The score of each path is calculated by considering the minimum score of weak unification along each path (Equation 2). The tail which has been reached via the highest path score is then considered as the final answer.

*Training.* The objective of the training is to minimise the following cross-entropy loss:

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\theta}} = - \quad & \textstyle\sum_{t\in\mathcal{T}} \log(Pr(\boldsymbol{t}|\boldsymbol{Z})Pr(\boldsymbol{Z}|\boldsymbol{h_q}, r_q, \mathcal{G}; \boldsymbol{\theta})) \\ & - \textstyle\sum_{t\notin\mathcal{T}} \log(1 - Pr(\boldsymbol{t}|\boldsymbol{Z})Pr(\boldsymbol{Z}|\boldsymbol{h_q}, r_q, \mathcal{G}; \boldsymbol{\theta}))\end{aligned}$$

where $\mathcal{T}$ is the set of all tails in $\mathcal{G}$ and $\boldsymbol{\theta}$ denotes all the parameters of the proposed model. During training the relation prediction module and embedding representation of all nodes in $\mathcal{G}$ are learnt. The relations are inferred by mapping the corresponding embedding to the nearest relation representation.

In order to apply our proposed model for out-of-domain context queries $r_q(C_q, X)$, e.g. Commonsense Question Answering, where the context is distributionally different from the nodes of CKG, the model dynamically leverages a generative SEQ2SEQ model in addition to the CKG. For this purpose, a narrative which describes a commonsense situation $\boldsymbol{C}$, a question $\boldsymbol{q}$ with respect to $\boldsymbol{C}$ and a set of possible answer candidates $\mathcal{A} = \{a^0, a^1, \ldots, a^m\}$ is provided, where the most plausible answer is required to be estimated by the model. To this end, by applying OpenIE6 (Kolluru et al. 2020), all the residing information in a given context is extracted, namely $\mathcal{I}_C = \{i_1, i_2, \ldots, i_n\}$. To increase the coverage of background knowledge, we use COSMO (Moghimifar et al.

2020), where for each of the extracted information in $\mathcal{I}_C$, a follow-up event is generated, which forms a set of triples, $\mathcal{G}_C$. The reasoning process is delivered similar to the above-mentioned explanation for each information in $\mathcal{I}_C$, by using both CKG and triples generated by the SEQ2SEQ model, and the final answer is estimated by:

$$\arg\max_{a \in \mathcal{A}} \max_{\boldsymbol{Z} \in \mathcal{Z}} \sum_{j=1}^{n} \log Pr(a|\boldsymbol{Z}) Pr(\boldsymbol{Z}|\boldsymbol{i}_j, r_q, \mathcal{G}, \mathcal{G}_C) \tag{7}$$

**Algorithm 1** Zero-Shot Commonsense Question Answering

**Input**: Context $\boldsymbol{C}$, Question $\boldsymbol{q}$, Answer Choices $\boldsymbol{A}$, Knowledge Graph $\mathcal{G}$
**Output**: The most plausible answer $a$
$\mathcal{I}_C \leftarrow$ Extract all information in $C$
Create an empty list $\mathcal{C}$
**for** $i$ in $\mathcal{I}_C$ **do**
    $c_k \leftarrow$ top-$k$ similar nodes to $i$ from $\mathcal{G}$
    Append all triples in $\mathcal{G}$ where $c_k$ is the head to $\mathcal{C}$
    $\mathcal{G}_C \leftarrow$ COSMO($i$)
    Append items in $\mathcal{G}_C$ to $\mathcal{C}$
**end for**
**for** $a$ in $\mathcal{A}$ **do**
    $score_a \leftarrow$ Solve Equation 7 to get the associated score with $a$
**end for**
Choose the answer $a$ with the highest score as the final answer

## 4. Experimental Results

In this section, we report the evaluation of our proposed model on the task of CKG completion and zero-shot CQA. For the latter task, we use the learnt logic rules from ATOMIC to perform multi-step reasoning. At each step of reasoning, we use SEQ2SEQ model proposed by (Moghimifar et al. 2020) to generate new facts based on the query, and perform reasoning over this constructed knowledge graph. The correct answer is then chosen by using the scoring function proposed by (Moghimifar et al. 2020) between entailed nodes and candidate answers.

### 4.1 Evaluation Metrics

For the task of CKG completion, followed by similar works on knowledge base completion task, we report the results in form of HITS and Mean Reciprocal Rank (MRR). For reporting the score of a gold target entity, the other valid entities are filtered out (Dettmers et al. 2018). The reported the results are the average score measured for queries $(h, r, ?)$ and $(t, r^{-1}, ?)$. For CQA we report the accuracy of the model in choosing the correct answer.

### 4.2 Baseline

For the task of CKG completion, we compare our proposed model to the state-of-the-art models in traditional KB completion and CKG completion. To this end we use DistMult (Yang et al. 2014), ComplEx (Trouillon et al. 2016), ConvE (Dettmers et al. 2018), RotatE (Sun et al. 2018), and Malaviya (Malaviya et al. 2020).

For the task of CQA, we compare our model to a pretrained language model, GPT (Radford et al. 2018, 2019), and the state-of-the-art CQA models, COMET (Bosselut and Choi 2019) and CosMo (Moghimifar et al. 2020). We also report the results of best performing supervised methods BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019).

### 4.3 Experimental Details

To train our model, each triple in form $r(h, t)$ in train set was converted to $r^{-1}(t, h)$, to account for reverse relations as well. We have used the embedding size of 1024 for both node and relation embedding layer. To embed the nodes in CKGs, we have fine-tuned uncased BERT-Large (Devlin et al. 2019) for the objective of masked language model. For this purpose, a node is converted into $[CLS] + n_i + [SEP]$ and fed into BERT. The representation of the token $[CLS]$ from last year of BERT is then used as node $n_i$ embedded representation. We used the maximum sequence of 128, and batch size of 64. Our relation predication module consists of two Linear layer. For all non-linearities in our model we have used ReLU. For optimisation purpose, SGD has been used, with staring learning rate of $10e - 4$, and decay rate of 0.9, if the loss of development set does not decrease after each epoch. We set the maximum depth of three for reasoning process. We have trained the model for 100 epochs.

Followed by (Malaviya et al. 2020), we have trained all the baselines for 200 epochs. During training the models were evaluated on development set, every 10 and 30 epochs, for ConceptNet-100K and ATOMIC, respectively. The checkpoint with the highest MRR was then selected for testing.

### 4.4 Datasets

- **ATOMIC** [1] is a social CKG, which contains facts about everyday social situations in form of *if-then* relations. This CKG contains more than 300K different entities, where they form more than 877K facts. (Sap et al. 2019a)
- **ConceptNet-100K** [2] contains general commonsense information about entities in form of triples. This dataset is a subset of ConceptNet 5 (Speer, Chin, and Havasi 2017). Over 80K entities in this CKG with 34 different relations form more than 100K facts.
- **SocialIQA** [3] contains commonsense questions about social situations. Each data in this dataset consists of a context, a question regarding that context and three possible candidate answers. In zero-shot setting, we report the results on test set and development set of this dataset, where each contain 1,954 and 2,224 questions, respectively.

[1] https://homes.cs.washington.edu/ msap/atomic/
[2] https://ttic.uchicago.edu/ kgimpel/commonsense.html
[3] https://maartensap.github.io/social-iqa/

| | ConceptNet-100k | | | | ATOMIC | | | |
|---|---|---|---|---|---|---|---|---|
| **Model** | MRR | HITS@1 | HITS@3 | HITS@10 | MRR | HITS@1 | HITS@3 | HITS@10 |
| DistMult | 8.97 | 4.51 | 9.76 | 17.44 | 12.39 | 9.24 | 15.18 | 18.30 |
| ComplEx | 11.40 | 7.42 | 12.45 | 19.01 | 14.24 | 13.27 | 14.13 | 15.96 |
| ConvE | 20.88 | 13.97 | 22.91 | 34.02 | 10.07 | 8.24 | 10.29 | 13.37 |
| RotatE | 19.89 | 14.45 | 25.32 | 37.56 | 10.61 | 8.56 | 10.76 | 14.98 |
| Malaviya et al. | 51.11 | 39.42 | 59.58 | 73.59 | 10.33 | 8.41 | 10.79 | 13.86 |
| Ours (Linear) | 61.16 | 55.29 | 62.62 | 73.37 | 51.05 | 45.21 | 50.11 | 64.18 |
| **Ours (GCN)** | **74.29** | **71.04** | **75.58** | **80.33** | **52.56** | **45.25** | **50.25** | **73.31** |

Table 2: Results on CKG completion task, on ConceptNet-100K and ATOMIC.

### 4.5 Results

Table 2 summarises the performance of our model in comparison with the baselines, on ATOMIC and ConceptNet-100K. Our model outperforms all of the baselines on ConceptNet-100k in all metrics. It can be seen that the improvement over the best baseline on MRR is around 13%. The closeness of HIT@1 results to MRR suggests that on this dataset, our model can estimate the tail of a relation in the first prediction with high probability. On ATOMIC the performance of our model outweighs all of the baselines in all metrics as well. However, the improvement over the best baseline is less than the performance on ConceptNet-100k. This is mostly due to the more challenging nature of ATOMIC, where the graph is more sparse and larger (Malaviya et al. 2020).

To further evaluate the performance of our model on new/unseen nodes, we choose a subset of the test set of ATOMIC and ConceptNet-100K, where for any $(h, r, t)$ either $h$ or $t$ is not seen by the model in the train set. Table 2 summarises the results of the conducted experiment on ConceptNet-100K and ATOMIC. On ConceptNet-100K our proposed model outperforms the baselines by up to 22 points on MRR. The gap between our model and the second best model decrease as we move from HITS@1 to HITS@10. This suggested that on contrary to the baselines our model performs better in estimating the probability of query with higher accuracy. On ATOMIC our model achieves a MRR of 46.41, which is 23 points higher than the second best model. As it can be seen from table 3, comparison of performance of different models on ConceptNet-100K and ATOMIC shows a noticeable drop in performance for models which rely on structural information of CKGs. This observation suggests that larger and sparser (lowest density) CKG are more challenging to reason over.

| | ConceptNet-100k | | | | ATOMIC | | | |
|---|---|---|---|---|---|---|---|---|
| **Model** | MRR | HITS@1 | HITS@3 | HITS@10 | MRR | HITS@1 | HITS@3 | HITS@10 |
| DistMult | 8.68 | 5.38 | 9.33 | 15.23 | 11.49 | 9.16 | 11.83 | 16.3 |
| ComplEx | 10.33 | 6.51 | 11.24 | 17.31 | 12.96 | 10.65 | 13.9 | 17.08 |
| ConvE | 16.55 | 10.19 | 18.79 | 28.08 | 9.04 | 7.05 | 9.42 | 12.74 |
| RotatE | 19.89 | 14.45 | 25.32 | 37.56 | 10.61 | 8.56 | 10.76 | 14.98 |
| Malaviya et al. | 43.60 | 39.33 | 49.41 | 66.58 | 23.43 | 20.54 | 24.1 | 27.43 |
| **Ours** | **65.72** | **57.49** | **61.7** | **71.46** | **46.41** | **43.31** | **45.94** | **47.24** |

Table 3: Results on CKG completion task, on unseen subset of ConceptNet-100K and ATOMIC.

Table 4 represents the performance of our model on the task of zero-shot commonsense question answering in comparison to the baselines. On unsupervised model, our model outperforms all the baselines on both test set and development set of SocialIQA. This improvement is mostly due to using learnt rules on ATOMIC to perform multi-hop reasoning, rather than exhausting all possible paths. However, the gap between unsupervised and supervised models suggest that the knowledge in ATOMIC may not be sufficient in answering all questions of SocialIQA.

| | MODEL | Dev Acc. | Test Acc. |
|---|---|---|---|
| Unsupervised | Random | 33.3 | 33.3 |
| | GPT | 41.8 | 41.7 |
| | GPT2-117M | 40.7 | 41.5 |
| | GPT2-345M | 41.5 | 42.5 |
| | GPT2-762M | 42.5 | 42.4 |
| | COMET-CA | 48.7 | 49.0 |
| | COMET-CGA | 49.6 | 51.9 |
| | CosMo | 54.8 | 55.0 |
| | **Ours** | **55.1** | **55.4** |
| Supervised | BERT-Large | 66.0 | 66.4 |
| | RoBERTa | 76.6 | 77.8 |
| | human | 86.9 | 84.4 |

Table 4: The accuracy of answer prediction of our proposed model compared to the state-of-the-art models on SocialIQA, on development and test set.

| **ATOMIC** |
|---|
| `xIntent(X,Y):-xIntent(X,Z),xIntent(Z,Y)` |
| `xNeed(X,Y):-xReact(Y,X)` |
| `xIntent(X,Y):-oWant(Y,X)` |
| **ConceptNet-100K** |
| `causes(X,Y):-causes(X,Z),causes(Z,Y)` |
| `isa(X,Y):-partof(X,Z),isa(Z,Y)` |
| `relatedto(X,Y):-relatedto(X,Z),relatedto(Z,Y)` |

Table 5: Examples of rules learned by our proposed relation prediction module.

Table 5 provides examples of generated rules by our model on ATOMIC and ConceptNet-100k. On ATOMIC, the first rule is based on transition, and the second and third rules are inverse rules. Similarly, on ConceptNet-100K the first and third rules are transitive, and the second rule is a compositional rule. All provided rules are diverse and meaningful, and can be used for explaining the inference process of our model. For instance, consider a query of *xIntent(Alex drives Jesse there, ?)*. Based on first rule from Table 3, `X` is unified by *Alex drives Jesse there*, and `Z` is unified by *Alex helps Jesse* (from triples of ATOMIC). Then, the query is updated to *xIntent(Alex helps Jesse, ?)* and `Y` is unified by *to be of assistance* (from triples of ATOMIC), hence the answer to query. The path generated by this example is *Alex drives Jesse there* $\xrightarrow{\text{xIntent}}$ *Alex helps Jesse* $\xrightarrow{\text{xIntent}}$ *to*

*be of assistance*. Therefore, two nodes are connected via a new link: *Alex drives Jesse there* $\xrightarrow{\text{xIntent}}$ *to be of assistance*.

Consider the following query from ConceptNet-100K, $HasProperty(novel, ?)$. Based on the relation of the query, our rule creator module can estimate the following rule:

```
HasProperty(X,Y):-IsA(X,Z),HasProperty(Z,Y)
```

According to this rule, `X` is unified by *novel*, and `Z` is unified by *book* (from triples of ConceptNet-100K). Then, the query is updated to *HasProperty(book, ?)* and `Y` is unified by *expensive* (from triples of ConceptNet-100K), resulting the answer to the query, by generating the following path: *novel* $\xrightarrow{\text{IsA}}$ *book* $\xrightarrow{\text{HasProperty}}$ *expensive*, hence *novel* $\xrightarrow{\text{HasProperty}}$ *expensive*.

## 5. Conclusion

In this work, we present a dynamic neural-symbolic reasoner, based on weak unification and forward chaining. The proposed reasoner leverages characteristics of graph and node embeddings to learn rules for multi-step reasoning on Commonsense Knowledge Graphs (CKGs). This process helps generalising the inference to unseen events. We showed that our model outperforms state-of-the-art models on both tasks of CKG completion and zero-shot Commonsense Question Answering.

## References


Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*. Springer, pages 722–735.

Banerjee, Pratyay and Chitta Baral. 2020. Self-supervised knowledge triplet learning for zero-shot question answering. *arXiv preprint arXiv:2005.00316*.

Bosselut, Antoine and Yejin Choi. 2019. Dynamic knowledge graph construction for zero-shot commonsense question answering. *arXiv e-prints*.

Bosselut, Antoine, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for knowledge graph construction. In *Association for Computational Linguistics (ACL)*.

Carlson, Andrew, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka, and Tom M Mitchell. 2010. Toward an architecture for never-ending language learning. In *Twenty-Fourth AAAI conference on artificial intelligence*.

Davis, Ernest and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103.

Dettmers, Tim, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*.

Ehrlinger, Lisa and Wolfram Wöß. 2016. Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, 48(1-4):2.

Fellbaum, Christiane. 2010. Wordnet. In *Theory and applications of ontology: computer applications*. Springer, pages 231–243.

Hwang, Jena D, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2020. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. *arXiv preprint arXiv:2010.05953*.

Ji, Shaoxiong, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*.

Johnson, Jeff, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*.

Kolluru, Keshav, Vaibhav Adlakha, Samarth Aggarwal, Soumen Chakrabarti, et al. 2020. Openie6: Iterative grid labeling and coordination analysis for open information extraction. *arXiv preprint arXiv:2010.03147*.

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

Malaviya, Chaitanya, Chandra Bhagavatula, Antoine Bosselut, and Yejin Choi. 2020. Commonsense knowledge base completion with structural and semantic context.

Moghimifar, Farhad, Lizhen Qu, Terry Yue Zhuo, Gholamreza Haffari, and Mahsa Baktashmotlagh. 2021a. Neural-symbolic commonsense reasoner with relation predictors. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 797–802, Association for Computational Linguistics.

Moghimifar, Farhad, Lizhen Qu, Yue Zhuo, Mahsa Baktashmotlagh, and Gholamreza Haffari. 2020. Cosmo: Conditional seq2seq-based mixture model for zero-shot commonsense question answering. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Moghimifar, Farhad, Lizhen Qu, Yue Zhuo, Gholamreza Haffari, and Mahsa Baktashmotlagh. 2021b. Neural-symbolic commonsense reasoner with relation predictors. *arXiv preprint arXiv:2105.06717*.

Mostafazadeh, Nasrin, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. Glucose: Generalized and contextualized story explanations. *arXiv preprint arXiv:2009.07758*.

Paulheim, Heiko. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, 8(3):489–508.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *Unpublished ms. available through a link at https://blog.openai.com/language-unsupervised/.*

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Sap, Maarten, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019a. Atomic: an atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Sap, Maarten, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473.

Sap, Maarten, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*.

Sessa, Maria I. 2002. Approximate reasoning by similarity-based sld resolution. *Theoretical computer science*.

Shang, Chao, Yun Tang, Jing Huang, Jinbo Bi, Xiaodong He, and Bowen Zhou. 2019. End-to-end structure-aware convolutional networks for knowledge base completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3060–3067.

Shwartz, Vered, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629.

Speer, Robyn, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Suchanek, Fabian M, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706.

Sun, Zhiqing, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2018. Rotate: Knowledge graph embedding by relational rotation in complex space. In *International Conference on Learning Representations*.

Toutanova, Kristina and Danqi Chen. 2015. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd workshop on continuous vector space models and their compositionality*, pages 57–66.

Trouillon, Théo, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*, pages 2071–2080.

Wang, Bin, Guangtao Wang, Jing Huang, Jiaxuan You, Jure Leskovec, and C-C Jay Kuo. 2021. Inductive learning on commonsense knowledge graph completion. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, IEEE.

Wang, Quan, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743.

Wu, Wentao, Hongsong Li, Haixun Wang, and Kenny Q Zhu. 2012. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 481–492, ACM.

Yang, Bishan, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*.

Zellers, Rowan, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Empirical Methods in Natural Language Processing*.

Zhang, Hongming, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020. Aser: A large-scale eventuality knowledge graph. In *Proceedings of The Web Conference 2020*, pages 201–211.