Learning from Missing Relations: Contrastive Learning with Commonsense Knowledge Graphs for Commonsense Inference

Anonymous ACL submission

Abstract

Commonsense inference poses a unique challenge to reason and generate the physical, social, and causal conditions of a given event. Existing approaches to commonsense inference utilize commonsense transformers, which are large-scale language models that learn commonsense knowledge graphs. However, they suffer from a lack of coverage and expressive diversity of the graphs, resulting in a degradation of the representation quality. In this paper, we focus on addressing missing relations in commonsense knowledge graphs, and propose a novel contrastive learning framework called SOLAR\(^1\). Our framework contrasts sets of semantically similar and dissimilar events, learning richer inferential knowledge compared to existing approaches. Empirical results demonstrate the efficacy of SOLAR in commonsense inference of diverse commonsense knowledge graphs. Specifically, SOLAR outperforms the state-of-the-art commonsense transformer on commonsense inference with ConceptNet by 1.84% on average among 8 automatic evaluation metrics. In-depth analysis of SOLAR sheds light on the effects of the missing relations utilized in learning commonsense knowledge graphs.

1 Introduction

Commonsense inference, reasoning of unobserved conditions from an observed event, is an important but challenging task in natural language processing (NLP) (Rashkin et al., 2018; Bosselut et al., 2019; Yuan et al., 2020; Hwang et al., 2021). This is easy for humans, but still out of the reach of current artificial intelligence systems. Commonsense inference aims to generate textual descriptions of the inference results, which is more in line with the process of humans reasoning based on their knowledge. For a given event “X walks into a hospital”, the causal conditions (e.g., what to do before and after the event), physical conditions (e.g., capability and location of entities), and social conditions (the intention and reaction of X) of the event are to be inferred.

Recent studies on commonsense inference have adopted commonsense transformers (Bosselut et al., 2019), which are large-scale language models trained on commonsense knowledge graphs (KGs) like ATOMIC (Sap et al., 2019) and ConceptNet (Speer et al., 2017). Such models are grounded on the hypothesis that language models can memorize facts in their parameters during training (Petroni et al., 2019; Roberts et al., 2020). Despite these efforts, commonsense transformer models still suffer from two main obstacles inherent in commonsense KGs: (1) lack of coverage and (2) expressive diversity of the graphs. First, commonsense KGs lack the coverage required to be applicable for diverse

\(^1\)Code available at [https://anonymous.4open.science/r/solar-commonsense_inference-3787](https://anonymous.4open.science/r/solar-commonsense_inference-3787)
situations in the real world (Li et al., 2016; Saito et al., 2018). In ATOMIC, even with the possibility of far more commonsense properties being relevant, any single event has only 2.2 commonsense properties directly related on average. Second, with the non-canonical and free-form text representation for the nodes in commonsense KGs, semantically identical or similar expressions of events are represented as distinct nodes (Malaviya et al., 2020). For example, “PersonX is fond of dogs” and “PersonX likes dogs” are semantically identical, but represented as distinct nodes. The expressive diversity makes commonsense KGs substantially sparser than conventional KGs. Owing to the lack of coverage and expressive diversity, significant amount of relations are missing in commonsense KGs.

In this study, we focus on learning from missing relations in commonsense KGs for commonsense inference. Our key observation is that semantically identical or similar events can have the same relations as shown in Figure 1. For example, “PersonX likes dogs” and “PersonX loves animals” are semantically similar to “PersonX loves dogs”, and the inference that “PersonX wants to adopt one” can be drawn from any of those events. Modeling such missing relations helps the model learn richer representations from commonsense KGs. Current approaches for alleviating the sparsity of commonsense KGs, such as automatic commonsense KG completion (Li et al., 2016; Saito et al., 2018; Malaviya et al., 2020), do not effectively address missing relations, because they consider only existing relations as valid. Therefore, this problem remains unexplored.

We propose a novel learning framework of commonsense transformers, called Self-supervised cOntrastive LeArning with missing Relations (SOLAR), to address the aforementioned problem. Our framework trains large-scale language models to learn both existing and missing relations with self-supervised contrastive learning. Specifically, we construct sets of examples including semantically similar events that can share relations based on the similarity of language representations. Each set contains semantically similar events within it, while events from other sets are semantically dissimilar. We then contrast each set of examples with the other sets. This allows the model to identify the interrelationship between semantically similar events and their relations, leading to a better understanding of missing relations in commonsense KGs.

We evaluate our framework for commonsense inference on three commonsense KGs: ConceptNet (Speer et al., 2017), ATOMIC (Sap et al., 2019), and ATOMIC$_{20}$ (Hwang et al., 2021). Empirical results show that SOLAR outperforms the state-of-the-art commonsense transformers on commonsense inference. In particular, for ConceptNet, SOLAR with BART-large (Lewis et al., 2020) outperforms COMET (Hwang et al., 2021) with BART-large by 1.84% on average among 8 automatic evaluation metrics. In addition, we observe that SOLAR with BART-base produces comparable results to COMET with BART-large, which validates that our framework is superior to existing approaches in terms of both effectiveness and efficiency. Our main contributions are as follows:

- We present a novel contrastive learning framework for commonsense transformers, called SOLAR, that learns from both existing and missing relations in commonsense KGs.
- We develop a principled scheme for constructing positive and negative sets of examples with commonsense KGs based on similarities of events in language representations.
- We verify that SOLAR establishes new state-of-the-art results in commonsense inference across diverse commonsense KGs.

2 Related Work

2.1 Commonsense Inference

In NLP domain, several studies have proposed commonsense inference models that utilize commonsense KGs. Rashkin et al. (2018) proposed Event2Mind, a commonsense KG that involves a textual description of a person’s response or intention of daily events. Sap et al. (2019) proposed ATOMIC knowledge graph as an extension of Event2Mind with more relations and tuples. Both studies trained on the GRU model based on their proposed graph to learn commonsense inference. Moreover, recent studies have shown that pre-trained language models store various types of fact knowledge in their latent parameters (Petroni et al., 2019; Roberts et al., 2020).Bosselut et al. (2019) revealed that language models can directly express commonsense knowledge by training them on commonsense KGs. Hwang et al. (2021) showed that
KGs must be designed to contain knowledge that is not already expressible by language models. Gabriel et al. (2021) focused on discourse-level commonsense inference, and Yuan et al. (2020) proposed a language model architecture for logically consistent commonsense reasoning. Previous studies have proposed training language models on existing tuples in commonsense KGs for commonsense inference. In our work, we focus on addressing the missing relations of commonsense KGs for better commonsense inference.

2.2 Contrastive Learning

Contrastive learning has shown promising performances in computer vision (Henaff, 2020; He et al., 2020). SimCLR (Chen et al., 2020b) introduced a simple but powerful contrastive learning approach and showed a competitive performance with supervised learning approaches. Contrastive learning is also widely used in natural language processing, where a model obtains unsupervised representations by learning to predict positive or negative pairs. Mikolov et al. (2013) proposed an efficient method for learning word representations by classifying whether given words appear in the same context or not. Furthermore, contrastive learning has been adopted to improve the representations of pre-trained language models. Reimers and Gurevych (2019); Zhang et al. (2020b); Yan et al. (2021) introduced contrastive learning frameworks for enhancing the sentence representations. Lee et al. (2020) proposed a contrastive learning method to mitigate the exposure bias problem. Inspired by these studies, we propose a novel contrastive learning framework for commonsense representation learning with commonsense KGs. With our proposed framework, the model learns inferential knowledge from both existing and missing relations.

3 Methodology

In this section, we describe the model architecture and training procedure of the proposed framework.

3.1 Notation

We define $G = (V, E)$ as the commonsense knowledge graph that consists of a set of nodes $V$ and a set of edges $E$. Following the notation from COMET (Bosselut et al., 2019), we denote each knowledge tuple from the knowledge graph as $\{s, r, o\}$, where $s$ is the phrase subject, $r$ is the relation, and $o$ is the phrase object of the tuple. Here, $s$ and $o$ are natural language sequences, and $r$ is a single special token (e.g., <xIntent>). Note that $s, o \in V$ and $\{s, r, o\} \in E$. We define $S$ as the set of all existing subjects from the knowledge graph, and it follows that $S \subset V$. Finally, we denote the generative language model to be trained as $f(\cdot)$ and a projection layer at the top of the model as $g(\cdot)$. 

Figure 2: Illustration of contrastive learning of commonsense tuples. (a) Based on adversarially sampled root subjects, semantically similar subjects are sampled. (b) Subjects and relation-object pairs connected to them are projected to separate hidden representations through a generative language model and a projection layer. (c) Hidden representations obtained from the same root subject are considered as positive pairs, and those obtained from other root subjects are considered as negative pairs for contrastive learning.
we use BERTScore (Zhang et al., 2020a) between

Algorithm 1 Set Construction Algorithm.

Input: root subjects $S_{root}$, number of root subjects $N$, edges $E$, set size $2m$, threshold $\delta$, BERTScore function $b(\cdot, \cdot)$, base model $f(\cdot)$, projection layer $g(\cdot)\\n$for $s_i \in S_{root}$ do
Initialize $G_i$ as $\emptyset$

for $j \in \{1, \ldots, m\}$ do
if $j = 1$ then $s_j \leftarrow s_i$
else repeat
$\triangleright$ Sample similar subject $s_j \leftarrow \text{sample}(S)$
until $b(f(s_j), f(s_i)) > \delta$
get tuple $\{s_j, r_j, o_j\} \in E$ containing $s_j$
$z_{2j-1}^i \leftarrow g(f(s_j))$
$z_{2j}^i \leftarrow g(f(r_j \oplus o_j))$
$G_i \leftarrow G_i \cup \{z_{2j-1}^i, z_{2j}^i\}$
end if
end for
return $G_1, G_2, \ldots, G_N$

3.2 Commonsense Representation Learning

To improve commonsense representations of the language model prior to learning commonsense inference, we first proceed with commonsense representation learning through contrastive learning of commonsense tuples and commonsense reconstruction.

Contrastive learning of commonsense tuples. Inspired by our key observation that semantically identical or similar events can have same relations, we propose a novel commonsense representation learning method based on contrastive learning.

The overall procedure of the proposed method is depicted in Figure 2. First, we obtain a set of $N$ root subjects $S_{root} = \{s_1, s_2, \ldots, s_N\}$ through adversarial sampling on $S$. The adversarial sampling procedure is designed such that pairwise semantic similarity of subjects in $S_{root}$ lies between minimum similarity $\alpha$ and maximum similarity $\beta$. Here, we use BERTScore (Zhang et al., 2020a) between phrase subjects as the semantic similarity metric.

We then obtain positive and negative pairs by constructing $N$ sets $G_1, G_2, \ldots, G_N$ containing hidden representations, where each $G_i$ corresponds to a root subject $s_i \in S_{root}$. For an arbitrary element $s_i \in S_{root}$, we first sample $m$ tuples $\{s_j, r_j, o_j\} (j = 1, 2, \ldots, m)$ from $E$ that contain subjects $s_j$ semantically similar to $s_i$. Each $s_j$ and $r_j \oplus o_j$ is projected to hidden representations $z_{2j-1}^i = g(f(s_j))$ and $z_{2j}^i = g(f(r_j \oplus o_j))$, and added to $G_i$. Here, $\oplus$ denotes concatenation of two sequences. Repeating for $m$ times, the constructed set $G_i$ contains $2m$ hidden representations derived from subjects that are semantically similar to the root subject $s_i$, and the relation-object pairs connected to them. Algorithm 1 summarizes the construction procedure.

We consider samples from the same set as positive pairs, and those from different sets are negative pairs in contrastive learning. We use NT-Logistic (the normalized temperature-scaled logistic) objective function (Chen et al., 2020b) as our training objective to maximize the agreement between positive pairs while minimizing the agreement between negative pairs. The formal expression of our objective function is given by the following equations:

$$l_{i}^{pos} = -\frac{\sum_{p,q=1}^{2m} \log \sigma(z_i^T z_p^q / \tau)}{2m}, \quad (1)$$

$$l_{i}^{neg} = -\frac{\sum_{i<j \leq N} \sum_{p,q=1}^{2m} \log \sigma(-z_i^T z_j^q / \tau)}{m(N-1)}, \quad (2)$$

$$L_{cont} = \frac{1}{N} \sum_{i=1}^{N} (l_{i}^{pos} + l_{i}^{neg}), \quad (3)$$

where $l_{i}^{pos}$ is the loss function over positive pairs in set $G_i$, and $l_{i}^{neg}$ is the loss function over negative pairs among set $G_i$ and the other sets. In addition, $\tau$ denotes the temperature parameter for temperature scaling. The model is trained to minimize the final objective $L_{cont}$, which is the mean of $l_{i}^{pos}$ and $l_{i}^{neg}$ for all $i = 1, 2, \ldots, N$.

Commonsense reconstruction. To further improve the representation of a single tuple, we propose a commonsense reconstruction task inspired by Lewis et al. (2020), in which the model learns to reconstruct noisy tuples into their original form. More specifically, we noise a commonsense tuple $\{s, r, o\}$ by randomly choosing one of the three elements, masking the span of the chosen element, and shuffling the order of the tuple. The model is trained to reconstruct the original tuple from the noisy tuple. This task complements the contrastive learning method by training the model to better understand the commonsense tuple itself. The objective of the commonsense reconstruction task is
to minimize \( L_{\text{recon}} \), computed by cross-entropy between the decoder output and the original tuple.

The model learns commonsense representations through multitask learning on the two aforementioned tasks simultaneously. Therefore, the final objective function of our framework is to minimize the combined loss:

\[
L_{\text{rep}} = \omega L_{\text{cont}} + (1 - \omega) L_{\text{recon}}.
\]  

3.3 Fine-tuning on Commonsense KGs

After learning commonsense representations, we remove the projection head and fine-tune the model with commonsense KGs to learn commonsense inference. The model learns to generate a phrase object \( o \) given a concatenation of phrase subject \( s \) and relation \( r \). The objective function of the task is as follows:

\[
L_{\text{infer}} = -\sum_{i=0}^{|E|} \log P_b(o_i|s_i,r_i)
\]  

3.4 Language Model Architecture

While SOLAR is agnostic to its generative language model architecture, for our experiments, we use BART (Lewis et al., 2020) with its pretrained parameters as our base generative language model. BART is a transformer-based sequence-to-sequence language model with a bidirectional encoder and a left-to-right autoregressive decoder. For commonsense representation learning (Section 3.2), we add a projection layer that maps the BART decoder output representations to a space where contrastive loss is applied. The projection head is then removed for fine-tuning on commonsense KGs (Section 3.3).

4 Experiments

In this section, we demonstrate the efficacy of our framework by comparing the commonsense inference performances of SOLAR with those of the state-of-the-art commonsense transformers.

4.1 Dataset

Commonsense KGs are widely used for evaluating the commonsense inference capability by measuring the plausibility of the generated inferences given unobserved events or entities. Hwang et al. (2021) developed an adversarial splitting method for dividing training, validation, and test sets that prevent overlapping subjects of knowledge tuples between the sets. We utilize the splitting method to evaluate the inference capability of the model for unseen events or entities. We use three commonsense KGs in our experiments: ConceptNet (Speer et al., 2017), ATOMIC (Sap et al., 2019), and ATOMIC\(^{20}\) (Hwang et al., 2021).

ConceptNet is a general commonsense knowledge graph. We use a subset of the graph provided by Li et al. (2016), which involves 36 relations and 300K tuples. The subset is divided into 265K, 5K, and 30K tuples for training, validation, and testing, respectively.

ATOMIC is a social commonsense knowledge graph that involves 9 relations with 877K tuples. The split of ATOMIC includes 710K, 80K, and 87K tuples for training, validation, and testing, respectively.

ATOMIC\(^{20}\) is a recently proposed large-scale commonsense knowledge graph, which involves 23 commonsense dimensions and contains 1.33M tuples. It includes physical-entity, social-interaction, and event-centered commonsense. ATOMIC\(^{20}\) is split into 1.08M, 10K, and 15K tuples for training, validation, and testing, respectively.

4.2 Experimental Settings

Baseline. We use COMET (Bosselut et al., 2019), the state-of-the-art commonsense transformers in commonsense inference, as the baseline. We use the public HuggingFace (Wolf et al., 2019) implementation of pre-trained BART (Lewis et al., 2020) as a language model and train it using SOLAR and COMET for comparison. BART-base has 6 transformer layers for encoder and decoder each with a hidden size of 768, whereas BART-large has 12 transformer layers for encoder and decoder each with a hidden size of 1024. For fine-tuning, we empirically choose the best number of epochs, learning rate, and batch size among \{1, 3, 5, 7, 11, 13\}, \{1e-4, 1e-5, 1e-6\}, and \{16, 32, 64, 128\}, respectively, and use the Adam optimizer with \( \beta_1 = 0.9, \beta_2 = 0.999 \).

Training details of SOLAR. In contrastive learning of commonsense tuples, we extract \( n \in \{4, 8, 16, 32\} \) root subjects while maintaining the similarity (%) between subjects with a minimum of \( \alpha \in \{40, 50, 60\} \) and a maximum of \( \beta \in \{70, 80\} \). We then sample \( m \in \{8, 16, 32\} \) semantically similar subjects based on previously extracted subjects. We set the temperature parameter \( \tau \) to 0.1.

In reconstructive learning tasks, we corrupt tu-
Table 1: Evaluation results (%) of commonsense inference with base models.

<table>
<thead>
<tr>
<th></th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
<th>METEOR</th>
<th>ROUGE-L</th>
<th>CIDEr</th>
<th>BERTScore</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConceptNet</td>
<td>COMET-base</td>
<td>15.60</td>
<td>10.26</td>
<td>6.88</td>
<td>4.84</td>
<td>11.79</td>
<td>16.61</td>
<td>33.41</td>
</tr>
<tr>
<td></td>
<td>SOLAR-base</td>
<td>17.12</td>
<td>11.55</td>
<td>8.10</td>
<td>5.79</td>
<td>12.90</td>
<td>18.25</td>
<td>38.91</td>
</tr>
<tr>
<td>ATOMIC</td>
<td>COMET-base</td>
<td>53.03</td>
<td>33.97</td>
<td>23.13</td>
<td>16.90</td>
<td>34.05</td>
<td>56.07</td>
<td>74.63</td>
</tr>
<tr>
<td></td>
<td>SOLAR-base</td>
<td>53.59</td>
<td>34.51</td>
<td>23.89</td>
<td>17.82</td>
<td>34.24</td>
<td>56.60</td>
<td>75.24</td>
</tr>
<tr>
<td>ATOMIC-20</td>
<td>COMET-base</td>
<td>44.99</td>
<td>26.95</td>
<td>17.44</td>
<td>11.77</td>
<td>31.20</td>
<td>48.33</td>
<td>59.48</td>
</tr>
<tr>
<td></td>
<td>SOLAR-base</td>
<td>45.42</td>
<td>27.62</td>
<td>18.15</td>
<td>12.47</td>
<td>31.59</td>
<td>48.84</td>
<td>61.12</td>
</tr>
</tbody>
</table>

Table 2: Evaluation results (%) of commonsense inference with large models.

<table>
<thead>
<tr>
<th></th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
<th>METEOR</th>
<th>ROUGE-L</th>
<th>CIDEr</th>
<th>BERTScore</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConceptNet</td>
<td>COMET-large</td>
<td>17.88</td>
<td>11.35</td>
<td>7.13</td>
<td>4.00</td>
<td>13.47</td>
<td>19.36</td>
<td>37.72</td>
</tr>
<tr>
<td></td>
<td>SOLAR-large</td>
<td>19.28</td>
<td>12.73</td>
<td>8.57</td>
<td>5.62</td>
<td>14.69</td>
<td>20.89</td>
<td>43.15</td>
</tr>
<tr>
<td>ATOMIC</td>
<td>COMET-large</td>
<td>54.05</td>
<td>34.92</td>
<td>24.04</td>
<td>17.62</td>
<td>35.06</td>
<td>56.93</td>
<td>75.46</td>
</tr>
<tr>
<td></td>
<td>SOLAR-large</td>
<td>54.31</td>
<td>35.77</td>
<td>25.41</td>
<td>19.45</td>
<td>35.30</td>
<td>57.11</td>
<td>76.33</td>
</tr>
<tr>
<td>ATOMIC-20</td>
<td>COMET-large</td>
<td>46.08</td>
<td>28.23</td>
<td>18.70</td>
<td>12.86</td>
<td>32.22</td>
<td>49.44</td>
<td>62.13</td>
</tr>
<tr>
<td></td>
<td>SOLAR-large</td>
<td>46.51</td>
<td>28.99</td>
<td>19.52</td>
<td>13.73</td>
<td>32.53</td>
<td>49.76</td>
<td>63.24</td>
</tr>
</tbody>
</table>

Table 3: Ablation study of commonsense representation learning methods on ATOMIC-20.

<table>
<thead>
<tr>
<th>Cont. Recon.</th>
<th>BLEU-3</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓ ✓</td>
<td>18.27</td>
<td>61.15</td>
</tr>
<tr>
<td>✓ X</td>
<td>18.02</td>
<td>61.02</td>
</tr>
<tr>
<td>X ✓</td>
<td>17.89</td>
<td>60.90</td>
</tr>
<tr>
<td>X X</td>
<td>17.43</td>
<td>59.48</td>
</tr>
</tbody>
</table>

Figure 3: Validation loss of COMET-large and SOLAR-large on ATOMIC-20.

Metrics. To measure the commonsense inference capability of SOLAR, we use common evaluation metrics in the text generation: BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), CIDEr (Vedantam et al., 2015) and BERTScore (Zhang et al., 2020a).

4.3 Results

Overall performance. We evaluate SOLAR and COMET on three commonsense KGs and report the automatic evaluation results of generated inferences. In our result tables, we denote model names in form of (framework)-(BART model configuration). For example, SOLAR and COMET with BART-base are denoted by SOLAR-base and COMET-base, respectively.

Table 1 shows that SOLAR-base outperforms COMET-base for all KGs. By averaging over all metrics, SOLAR-base improves the performance of COMET-base on ConceptNet, ATOMIC, and ATOMIC-20 by 1.74%, 0.57%, and 0.65%, respectively. Experiments on large model configurations establish the new state-of-the-art results on commonsense inference with KGs. Table 2 shows that SOLAR-large outperforms COMET-large, the previous state-of-the-art, for all KGs and evaluation
Subject Relation Ground truth COMET SOLAR

PersonX is always busy xReact exhausted busy tired

sugar cube ObjectUse eat as food mix with sugar swetten coffee

PersonX gives PersonY a cup HinderedBy PersonY is not thirsty PersonX is allergic to water PersonX doesn’t have a cup

PersonX likes the movie HinderedBy They were too busy texting PersonX is allergic to the movie The movie is too boring

Table 4: Examples of commonsense inference from COMET and SOLAR in ATOMIC.

Analysis on commonsense inference. We provide further analysis on commonsense inference results of SOLAR and COMET. Figure 3 shows the validation loss curve for COMET-large and SOLAR-large. It is clearly observed that SOLAR gives smaller loss than COMET on validation sets, which indicates that SOLAR generalizes commonsense better than COMET. In addition, Table 4 shows examples of commonsense inference results by COMET and SOLAR. It can be observed that SOLAR generates plausible inferences with novel expressions, whereas COMET extracts words from the subject phrase to generate inferences, leading to trivial or wrong results. Another observation is that COMET is vulnerable to the annotation bias in KGs. For example, in ATOMIC, the word “allergic” frequently appears with relation “HinderedBy”, and COMET is biased to generate wrong inferences like “allergic to the movie”. In contrast, SOLAR makes better inference results without such bias.

Ablation study. We conduct an ablation study to measure the effectiveness of each component of our proposed framework. Table 3 shows that learning on both tasks performs better than learning on only one of the two tasks. We observe that contrastive learning of commonsense tuples is the key to our performance improvement that SOLAR achieves, and the reconstruction task also plays a role in the framework.

Acceptance of missing relations. We conduct a qualitative analysis of missing relations generated through our approach. Table 5 shows examples of tuple pairs and their similarity values measured by BERTScore. In the first row, “PersonX throws a huge party” and “PersonX throws a big party” are semantically similar, and each relation-object can be shared with the subject of the other (e.g., PersonX throws a huge party - oEffect - smile). In contrast, as in the last example, tuple pairs with a low similarity between subjects cannot share relation-object with one another. From these examples, we observe that tuple pairs with higher similarity between subjects generate more plausible tuples when their relation-object pair are shared, consistent with our intuition.

We further provide a quantitative analysis by measuring the acceptance rate of missing relations generated through our approach and comparing it with the overlap rate. Overlap rate is the probability of a missing relation already existing in the
<table>
<thead>
<tr>
<th>Similarity (%)</th>
<th>Subject</th>
<th>Relation – object</th>
<th>Plausible</th>
</tr>
</thead>
<tbody>
<tr>
<td>95.8</td>
<td>PersonX throws a huge party</td>
<td>oReact-important oEffect-smile</td>
<td>✓</td>
</tr>
<tr>
<td>95.3</td>
<td>handgun</td>
<td>AtLocation-army</td>
<td>✓</td>
</tr>
<tr>
<td>90.3</td>
<td>protective clothing safety gear</td>
<td>ObjectUse-keep them safe ObjectUse-protect from injury</td>
<td>✓</td>
</tr>
<tr>
<td>87.0</td>
<td>trash bags</td>
<td>ObjectUse-put things in ObjectUse-get rid of garbage</td>
<td>✓</td>
</tr>
<tr>
<td>82.0</td>
<td>PersonX takes PersonY to see a doctor</td>
<td>oEffect–get checked by doctor xWant-get dog checked</td>
<td>✗</td>
</tr>
<tr>
<td>70.1</td>
<td>PersonX hugs PersonY back</td>
<td>oReact-loved and needed oEffect-sweats in terror</td>
<td>✗</td>
</tr>
</tbody>
</table>

Table 5: Qualitative analysis on examples of similarity-based tuple extraction from ATOMIC. Similarity is measured by BERTScore between the subjects of tuples. Humans evaluate whether the tuples are plausible after the relation-objects are replaced by that of each other.

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU-3</th>
<th>CIDEr</th>
<th>BERTScore</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>17.44</td>
<td>59.48</td>
<td>63.11</td>
</tr>
<tr>
<td>Fine-tuning</td>
<td>17.38</td>
<td>59.11</td>
<td>63.08</td>
</tr>
<tr>
<td>Contrastive Learning</td>
<td><strong>18.15</strong></td>
<td><strong>61.12</strong></td>
<td><strong>63.27</strong></td>
</tr>
</tbody>
</table>

Table 6: Evaluation results of methods for learning from missing relations.

graph. To measure the acceptance rate of missing relations, we randomly sample 20 missing relations per similarity interval (total 120 samples) and ask human annotators to determine their plausibility. Three workers annotated each missing relation as accept if it is plausible or reject otherwise, and we used majority voting as the final annotation. Figure 4 shows the acceptance rate of the missing relations regarding semantic similarity of subjects. It shows that the acceptance rate of missing relation is proportional to the similarity, and if the tuples have a similarity of greater than 90%, then 90% of the missing tuples are then valid. In contrast, when the similarity dropped below 85%, the acceptance rate decreased drastically. The blue line in Figure 4 represents the overlap rate according to the similarity. For tuple pairs of high similarity exceeding 90%, the overlap rate is significantly lower (< 20%) than the acceptance rate, which shows that novel missing relations can be effectively identified through our method.

Methods for learning from missing relations. We investigate the effectiveness of our method for learning from missing relations. We compare our contrastive learning method with a fine-tuning method where missing relations are directly added to a commonsense KG and subsequently learned. We use missing relations generated on subjects with exceeding 90% similarity. Table 6 shows that our proposed contrastive learning method shows best performance, while fine-tuning method is worse than the baseline. We speculate that direct fine-tuning is vulnerable to unacceptable relations, while our proposed contrastive learning framework is robust to them. These results indicate that directly learning from missing tuples harm the commonsense inference capability of the model. We speculate that our approach can handle noise or incorrect missing relations by implicitly learning from missing relations.

5 Conclusion

We have presented a novel contrastive learning framework of commonsense transformers, called SOLAR, to effectively learn from missing relations in commonsense KGs. Moreover, we have developed a new construction scheme for positive and negative sets of examples based on similarities in language model representations. By utilizing our carefully designed methods, SOLAR effectively learns both existing and missing relations of events, alleviating the difficulties in learning commonsense KGs. Our empirical evaluations of diverse commonsense KGs demonstrate the efficacy of SOLAR in commonsense inference. In particular, SOLAR consistently outperforms the state-of-the-art commonsense transformers across all the evaluation metrics and commonsense KGs.
References


