# 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.

[1]Code available at https://anonymous.4open.science/r/solar-commonsense_inference-37E7

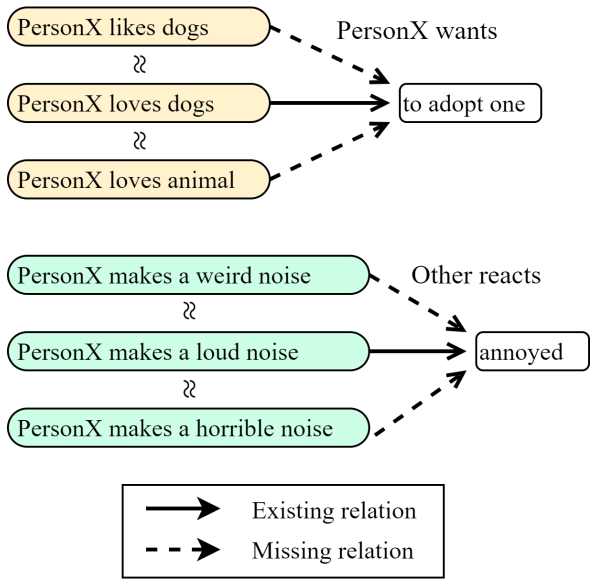


Figure 1: Illustration of missing relations of semantically similar events in commonsense KGs.

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 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 ***S**elf-supervised **cO**ntrastive **LeA**rning with missing **R**elations* (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$^{20}_{20}$ (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 pretrained 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

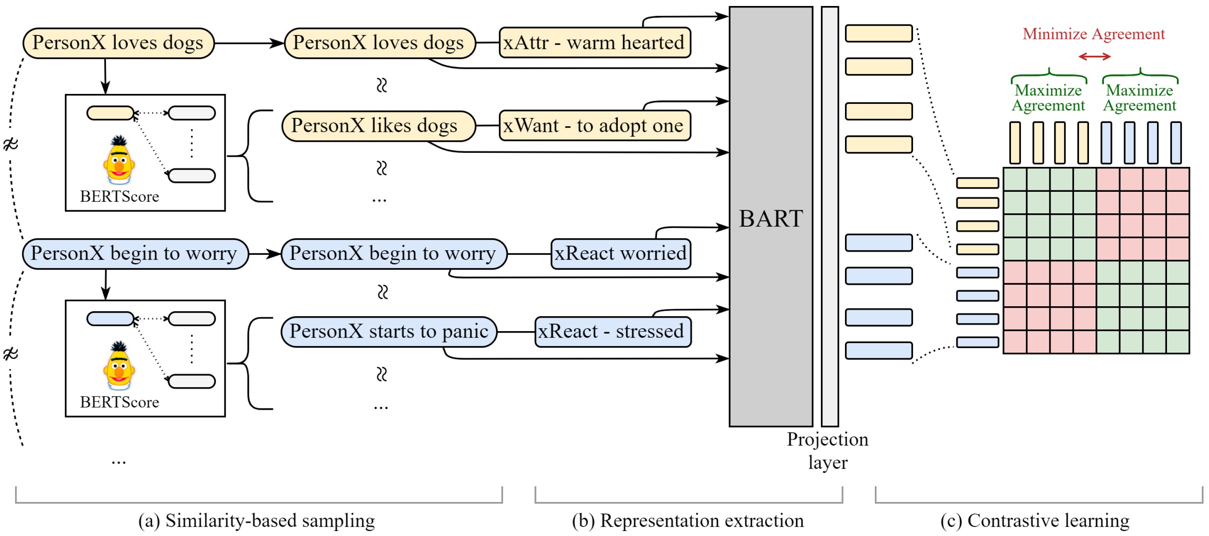


Figure 2: Illustration of contrastive learning of commonsense tuples. (a) Based on adversarially sampled root subjects, semantically similar subjects are sampled. (b) Subjects and relation-object pairs connected to them are projected to separate hidden representations through a generative language model 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.

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 $g(\cdot)$.

**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 $g(\cdot)$

```
for s_i ∈ S_root do
    Initialize G_i as ∅
    for j ∈ {1, ..., m} do
        if j = 1 then
            s^i_j ← s_i
        else
            repeat                          ▷ Sample similar subject
                s^i_j ← sample(S)
            until b(f(s^i_j), f(s_i)) > δ
        end if
        get tuple {s^i_j, r^i_j, o^i_j} ∈ E containing s^i_j
        z^i_{2j−1} ← g(f(s^i_j))
        z^i_{2j} ← g(f(r^i_j ⊕ o^i_j))
        G_i ← G_i ∪ {z^i_{2j−1}, z^i_{2j}}
    end for
end for
return G_1, G_2, ..., 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 $S_{root} = \{s_1, s_2, ..., s_N\}$ 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 $G_1, G_2, ..., G_N$ 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 $\{s_j, r_j, o_j\}$ $(j = 1, 2, ..., m)$ 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 $z^i_{2j-1} = g(f(s_j))$ and $z^i_{2j} = g(f(r_j \oplus o_j))$, 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:

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

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

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

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 $i = 1, 2, ..., N$.

**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 $L_{recon}$ 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:

$$L_{rep} = \omega L_{cont} + (1 - \omega) L_{recon}. \quad (4)$$

### 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:

$$L_{infer} = -\sum_{i=0}^{|E|} \log P_{\theta}(o_i | s_i, r_i) \quad (5)$$

### 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 pre-trained 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$^{20}_{20}$ (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$^{20}_{20}$** 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$^{20}_{20}$ 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-

| | | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-L | CIDEr | BERTScore |
|---|---|---|---|---|---|---|---|---|---|
| ConceptNet | COMET-base | 15.60 | 10.26 | 6.88 | 4.84 | 11.79 | 16.61 | 33.41 | 53.18 |
| | SOLAR-base | **17.12** | **11.55** | **8.10** | **5.79** | **12.90** | **18.25** | **38.91** | **53.86** |
| ATOMIC | COMET-base | 53.03 | 33.97 | 23.13 | 16.90 | 34.05 | 56.07 | 74.63 | 64.57 |
| | SOLAR-base | **53.59** | **34.51** | **23.89** | **17.82** | **34.42** | **56.60** | **75.24** | **64.78** |
| ATOMIC$^{20}_{20}$ | COMET-base | 44.99 | 26.95 | 17.44 | 11.77 | 31.20 | 48.33 | 59.48 | 63.11 |
| | SOLAR-base | **45.42** | **27.62** | **18.15** | **12.47** | **31.59** | **48.84** | **61.12** | **63.27** |

Table 1: Evaluation results (%) of commonsense inference with base models.

| | | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-L | CIDEr | BERTScore |
|---|---|---|---|---|---|---|---|---|---|
| ConceptNet | COMET-large | 17.88 | 11.35 | 7.13 | 4.00 | 13.47 | 19.36 | 37.72 | 54.07 |
| | SOLAR-large | **19.28** | **12.73** | **8.57** | **5.62** | **14.69** | **20.89** | **43.15** | **54.71** |
| ATOMIC | COMET-large | 54.05 | 34.92 | 24.04 | 17.62 | 35.06 | 56.93 | 75.46 | 64.84 |
| | SOLAR-large | **54.31** | **35.77** | **25.41** | **19.45** | **35.30** | **57.11** | **76.33** | **64.91** |
| ATOMIC$^{20}_{20}$ | COMET-large | 46.08 | 28.23 | 18.70 | 12.86 | 32.22 | 49.44 | 62.13 | 63.52 |
| | SOLAR-large | **46.51** | **28.99** | **19.52** | **13.73** | **32.53** | **49.76** | **63.24** | **63.58** |

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

| | Cont. | Recon. | BLEU-3 | CIDEr |
|---|---|---|---|---|
| SOLAR-base | ✓ | ✓ | **18.27** | **61.15** |
| | ✓ | ✗ | 18.02 | 61.02 |
| | ✗ | ✓ | 17.89 | 60.90 |
| | ✗ | ✗ | 17.43 | 59.48 |

Table 3: Ablation study of commonsense representation learning methods on ATOMIC$^{20}_{20}$

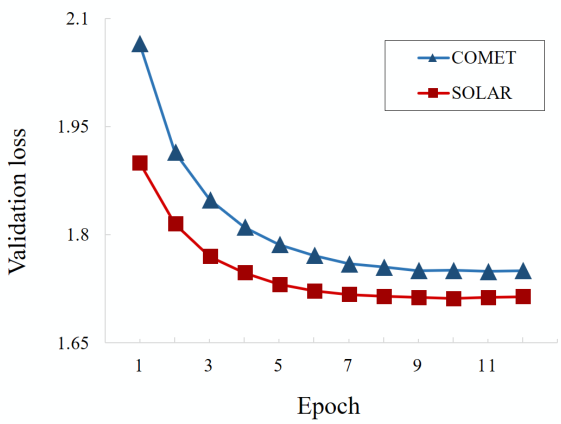


Figure 3: Validation loss of COMET-large and SOLAR-large on ATOMIC$^{20}_{20}$

ples by masking the span of each tuple elements and randomly shuffling the order. The span length is drawn from a Poisson distribution ($\lambda = 3$). SOLAR learns commonsense representation through multi-task approach, and we set the task weight as $\omega = 0.8$. In addition, we optimize the model using the RecAdam (Chen et al., 2020a) optimizer to prevent catastrophic forgetting during commonsense representation learning. We set the hyperparameters of the optimizer to $k = 0.001$ and $t_0 = 1000$. After representation learning, we set the same hyperparameters as the baseline. We report the best results among possible hyperparameter settings.

**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$^{20}_{20}$ 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$^{20}_{20}$.

metrics. We observe 1.84%, 0.70%, and 0.58% average performance improvement on ConceptNet, ATOMIC, and ATOMIC$^{20}_{20}$ respectively. Furthermore, SOLAR-base performs comparably to COMET-large on ATOMIC and ATOMIC$^{20}_{20}$, and performs better on ConceptNet, despite using only one-third of parameters. This shows the parameter-efficiency of our approach compared to COMET.

**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$^{20}_{20}$, the word "allergic" frequently appears with relation "HinderedBy", and COMET is biased to generate wrong inferences like "allergic to the moive". 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.

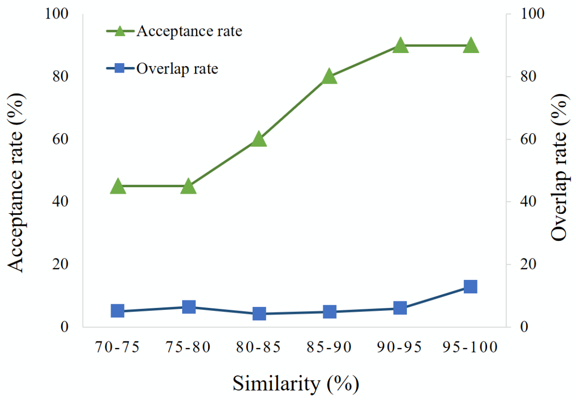


Figure 4: Acceptance and overlap rates of generated missing relations. Similarity is measured by BERTScore.

**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

| Similarity (%) | Subject | Relation – object | Plausible |
|---|---|---|---|
| 95.8 | PersonX throws a huge party<br>PersonX throws a big party | oReact-important<br>oEffect-smile | ✓ |
| 95.3 | handgun<br>pistol | AtLocation-army<br>AtLocation-pants | ✓ |
| 90.3 | protective clothing<br>safety gear | ObjectUse-keep them safe<br>ObjectUse-protect from injury | ✓ |
| 87.0 | trash bags<br>trashbins | ObjectUse-put things in<br>ObjectUse-get rid of garbage | ✓ |
| 82.0 | PersonX takes PersonY to see a doctor<br>PersonX takes PersonY to the vet | oEffect–get checked by doctor<br>xWant-get dog checked | ✗ |
| 70.1 | PersonX hugs PersonY back<br>PersonX screams at PersonY | oReact-loved and needed<br>oEffect-sweats in terror | ✗ |

Table 5: Qualitative analysis on examples of similarity-based tuple extraction from ATOMIC$^{20}_{20}$. 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.

| Method | BLEU-3 | CIDEr | BERTScore |
|---|---|---|---|
| Baseline | 17.44 | 59.48 | 63.11 |
| Fine-tuning | 17.38 | 59.11 | 63.08 |
| Contrastive Learning | **18.15** | **61.12** | **63.27** |

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


Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779.

Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. 2020a. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7870–7881.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020b. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR.

Saadia Gabriel, Chandra Bhagavatula, Vered Shwartz, Ronan Le Bras, Maxwell Forbes, and Yejin Choi. 2021. Paragraph-level commonsense transformers with recurrent memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12857–12865.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

Olivier Henaff. 2020. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6384–6392.

Seanie Lee, Dong Bok Lee, and Sung Ju Hwang. 2020. Contrastive learning with adversarial perturbations for conditional text generation. In *International Conference on Learning Representations*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Xiang Li, Aynaz Taheri, Lifu Tu, and Kevin Gimpel. 2016. Commonsense knowledge base completion. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1445–1455.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chaitanya Malaviya, Chandra Bhagavatula, Antoine Bosselut, and Yejin Choi. 2020. Commonsense knowledge base completion with structural and semantic context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2925–2933.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A Smith, and Yejin Choi. 2018. Event2mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 3982–3992.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426.

Itsumi Saito, Kyosuke Nishida, Hisako Asano, and Junji Tomita. 2018. Commonsense knowledge base completion and generation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 141–150.

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

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4444–4451.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075.

Chenxi Yuan, Chun Yuan, Yang Bai, and Ziran Li. 2020. Logic enhanced commonsense inference with chain transformer. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1763–1772.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020b. An unsupervised sentence embedding method by mutual information maximization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1601–1610.