Improved Contrastive Learning over Commonsense Knowledge Graphs for Unsupervised Reasoning

Rongwen Zhao and **Jeffrey Flanigan**
University of California, Santa Cruz
{rzhao17,jmflanig}@ucsc.edu

Abstract

Knowledge-augmented methods leverage external resources such as commonsense knowledge graphs (CSKGs) to improve downstream reasoning tasks. Recent work has explored contrastive learning over relation-aware sequence pairs derived from CSKG triples to inject commonsense knowledge into pre-trained language models (PLMs). However, existing approaches suffer from two key limitations: they rely solely on randomly sampled in-batch negatives, overlooking more informative hard negatives, and they ignore additional plausible positives that could strengthen training. Both factors limit the effectiveness of contrastive knowledge learning. In this paper, we propose an enhanced contrastive learning framework for CSKGs that integrates **hard negative sampling** and **positive set expansion**. Hard negatives are dynamically selected based on semantic similarity to ensure the model learns from challenging distinctions, while positive set expansion exploits the property that similar head entities often share overlapping tail entities, allowing the recovery of missing positives. We evaluate our method on unsupervised commonsense question answering and inductive CSKG completion using ConceptNet and ATOMIC. Experimental results demonstrate consistent improvements over strong baselines, confirming that our approach yields richer commonsense-aware representations and more effective knowledge injection into PLMs.

1 Introduction

Commonsense reasoning is fundamental for enabling machines to form assumptions about everyday situations and draw conclusions aligned with human understanding of commonly known facts (Davis and Marcus, 2015; Sap et al., 2020). Despite significant progress in natural language processing (NLP), endowing models with robust commonsense reasoning abilities remains an open challenge. This challenge has received growing attention in recent years with the release of versatile benchmark datasets targeting different aspects of commonsense reasoning. For example, Figure 1 illustrates a sample from the SocialIQA dataset (Sap et al., 2019b), which focuses on reasoning about human actions and their social implications. In parallel, the development of large-scale commonsense knowledge graphs (CSKGs), such as ConceptNet (Speer et al., 2017) and ATOMIC (Sap et al., 2019a), has motivated tasks like inductive CSKG completion to further test models' ability to generalize over unseen entities (Malaviya et al., 2020; Wang et al., 2021).

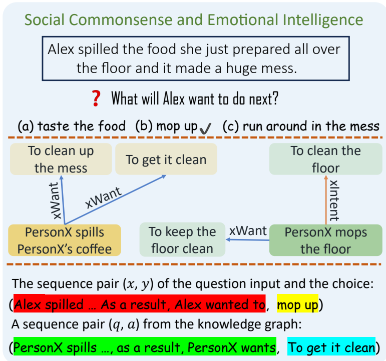

Figure 1: An example from a SocialIQA task focusing on reasoning about actions and social implications (**top**) (Sap et al., 2019b), with the relevant social commonsense knowledge triplets from ATOMIC (**middle**) (Sap et al., 2019a). The **bottom** shows a (input, choice) sequence pair of the example and a (premise, alternative) sequence pair of a knowledge graph triplet.

With the advent of large pre-trained language models (PLMs) (Devlin et al., 2018; Liu et al., 2019; Radford et al., 2019), fine-tuning PLMs on task-specific commonsense question answering (CSQA) datasets has led to strong results, in some cases approaching or surpassing human performance (He et al., 2020). However, reliance on large-scale human-annotated training data poses challenges, as such annotations are expensive and difficult to scale (Shwartz et al., 2020; Banerjee and Baral, 2020; Bosselut et al., 2021; Sun et al., 2022). Moreover, evidence shows that PLMs often

Proceedings of the First Workshop on Comparative Performance Evaluation: From Rules to Language Models associated with RANLP 2025, pages 165–178, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-102-8-017

exploit spurious correlations or shortcuts in data (Branco et al., 2021), rather than performing genuine commonsense reasoning or effectively leveraging external knowledge sources (Banerjee et al., 2021).

To mitigate these limitations, several unsupervised approaches based on CSKGs have been proposed. For instance, Ma et al. (2021); Kim et al. (2022) generate synthetic QA pairs from CSKG triples by treating the head entity with its relation as a query and the tail entity as the gold answer. Yet, the coverage of such methods is constrained by the incompleteness of CSKGs (Ju et al., 2022). More recently, Su et al. (2022) introduced a contrastive learning framework that pre-trains PLMs on (premise, alternative) pairs synthesized from CSKGs. While effective, this approach has two major shortcomings: (i) it relies on randomly sampled in-batch negatives, overlooking the importance of *hard negatives*, and (ii) it ignores potentially valuable positive examples inherent in CSKG structures. Both factors may limit the efficacy of the contrastive learning paradigm.

In this work, we propose an enhanced contrastive learning framework to better utilize CSKGs for commonsense knowledge representation. Our method incorporates two key components: **(i) hard negative sampling**, which dynamically selects informative negatives that are neither trivial nor indistinguishably similar, and **(ii) positive set expansion**, which leverages the property that similar head entities in CSKGs often share overlapping tail entities, thereby recovering missing positives. By integrating these mechanisms into the contrastive objective, we more effectively exploit the structure of CSKGs to improve knowledge injection into PLMs.

We evaluate our framework on two widely used CSKGs, ConceptNet and ATOMIC, across unsupervised CSQA benchmarks, including COPA (Roemmele et al., 2011), SIQA (Sap et al., 2019b) and CSQA (Talmor et al., 2019) and inductive CSKG completion tasks. Experimental results demonstrate consistent improvements over strong baselines, confirming that our framework generates superior commonsense-aware knowledge representations.

2 Preliminaries and Preprocessing

In this section, we first introduce some preliminaries used in this work. Then we will present the preprocessing details.

2.1 Task Definition

Our task is the following: given a common-sense knowledge graph $\mathcal{G}$ and a pre-trained language model $\mathcal{M}$, we construct a synthesized corpus of sequence pairs $\mathcal{D} = \{(p_1, a_1), ..., (p_i, a_i)\}$ from $\mathcal{G}$, where p is the head sequence and a is the natural language description of the tail entity. Then we further train $\mathcal{M}$ on the corpus $\mathcal{D}$ so $\mathcal{M}$ performs better on a given downstream commonsense-related task represented as $\mathcal{T} = \{(x_1, y_1), ..., (x_m, y_m)\}$ by encouraging M to generate superior commonsense-aware knowledge representation embeddings for the sequence pair (x_m, y_m). The corpus $\mathcal{D}$ is constructed from $\mathcal{G}$ using the method described in §2.3.

2.2 Notation

We define our commonsense knowledege graph $\mathcal{G}$ as a 4-tuple $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathcal{P})$, where the vertices are entities $\mathcal{E}$ and $\mathcal{R}$ are the set of relation types. $\mathcal{T}$ is the set of all edges, where each edge is a triple (h, r, t). $h \in \mathcal{E}$, $r \in \mathcal{R}$, and $t \in \mathcal{E}$ are the head entity, relation, and tail entity, respectively. $\mathcal{P}$ is the collection of all relations expressed in natural language, as shown in Appendix A.2. Additionally, following previous work (Ouyang et al., 2021; Su et al., 2022) we augment $\mathcal{G}$ with inverse edges: for each edge triple $(h, r, t) \in \mathcal{T}$ we add its reverse triple (h, r^{-1}, t) into $\mathcal{G}$.

2.3 Knowledge Graph Triple to Natural Language

In CSKGs, the entities h and t in $\mathcal{E}$ are in a free-form text format, and the relation r is a specific word or short phrase based on the corresponding CSKG. For example, (h, r, t) in ConceptNet could be (*Bottle, MadeOf, Plastic*) or (*PersonX spills PersonX's coffee, xWant, To get it clean*) in ATOMIC. . We use a set of templates for the relation r and its reverse relation r^{-1} in ATOMIC and ConceptNet. Following previous work (Hwang et al., 2021; Huang et al., 2021; Su et al., 2022), we first convert each edge triple (h, r, t) into a sequence pair (p, a) in natural language, consisting of a head sequence and its tail sequence. The original relation r is converted to the pre-defined natural language template and then connect it with the head entity h to form the head sequence p, while a is the natural language description of the tail entity t.

For example, in Figure 1, for the head node "*PersonX spills PersonX's coffee*", we concatenate it

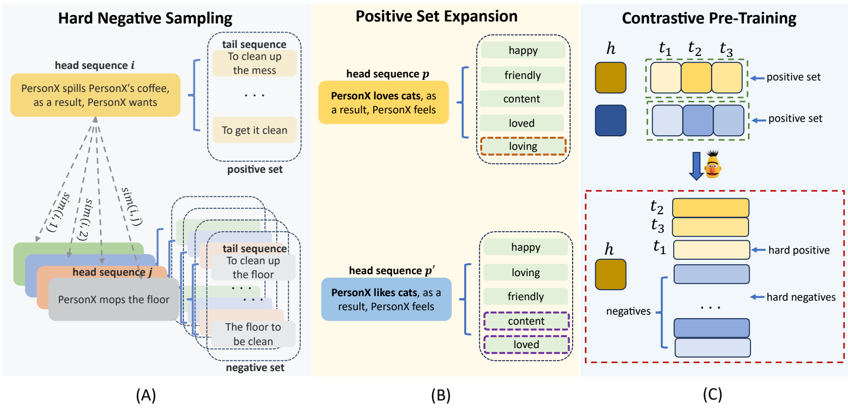

Figure 2: The steps in our contrastive learning framework. (A) **Hard Negative Sampling**: We dynamically sample hard negatives by the similarity of premise pairs. (B) **Positive Set Expansion**: We deliberately utilize the characteristic within the CSKGs that similar head entities are likely to share the same set of positive tail entities and expand the possible positive set mutually. (C) **Contrastive Training**: We integrate the updated sequence pairs into the existing multi-view contrastive learning framework to perform knowledge injection.

with the relation template of "xWant", resulting in the head sequence "*PersonX spills PersonX's coffee, as a result, PersonX wants.*" Similarly, for the reverse relation r^{-1}, we can also derive a sequence pair. Since for a head entity h, given a relation r, it may have n tail entities $\{t_1, t_2, ..., t_n\}$. Therefore, for a head sequence p, it may have a set of tail sequences $\{a_1, a_2, ..., a_n\}$.

2.4 Embedding Representation

After obtaining the sequence pair (p, a), we use a pre-trained language model (PLM) to get an initial embedding representation for the sequence pair. Specifically, for a sequence pair (p, a), where both p and a consist of sequence of tokens $\{x_0, ..., x_m\}$ and $\{y_0, ..., y_n\}$, respectively, We apply a PLM encoder to obtain the last hidden states of p and a, then use the hidden state of the first token, e_p and e_a as the embedding representation for p and a.

For a positive sequence pair (p, a), their representations in embedding space e_p and e_a should be close. We adopt the cosine similarity function to measure the distance of p and a:

$$sim(p, a) = cos(e_p, e_a)$$

3 Methodology

Our commonsense-aware knowledge representation learning framework, as shown in Figure 2, is divided into three steps: hard negative set sampling, positive set expansion, and contrastive knowledge fine-tuning. The input consists of a CSKG (e.g., ATOMIC) and a PLM (e.g., RoBERTa-Large). Given the synthesized CSKG sequence pairs obtained from §2.4, the goal is to inject the commonsense knowledge into the PLM by further training on the synthesized sequence pairs with enhanced contrastive learning.

We propose to enhance the existing contrastive learning framework for learning commonsense knowledge representation (Su et al., 2022). We propose two mechanisms to mitigate two issues that may impede the learning efficacy of the contrastive learning framework. First, we propose hard negative sampling to pay more attention to the hard ones instead of merely relying on random in-batch negatives (§3.1). Second, we propose to expand the positive set so that the missing positives could be recovered (§3.2). Finally, the PLM is trained with the adapted contrastive objective (§3.3).

3.1 Hard Negative Sampling

In this paper, we propose adapting the idea of hard negative sampling to the existing contrastive learning framework for the common sense-aware knowledge representation task. The learning framework learns commonsense knowledge representation with the contrastive information of the natural language sequence pairs. In particular, the existing method utilizes samples within the same mini-batch as negatives (Su et al., 2022), although such a strategy can significantly enhance training efficiency by repeatedly using the representations of in-batch negatives. However, this method ignores the difference of easy and hard negatives. Some literatures have theoretically and empirically proved that easy samples contribute less to the final learned representation (Bucher et al., 2016; Wu et al., 2017; Robinson et al., 2020; Zhang and Stratos, 2021). Recently, several adaptations in knowledge graph representation learning for knowledge graph completion and commonsense question answering also verify the importance of sampling hard negatives (Wang et al., 2022; Peng et al., 2022; Zhang and Li, 2022). The success of the contrastive representation benefits more from the hard ones, which means that the negatives that are difficult to distinguish are preferred instead of relying on randomly selected in-batch negatives.

To illustrate the proposed idea more precisely, consider the corpus $\mathcal{D}$ consisting of all triples converted from the CSKG $\mathcal{G}$ by the aforementioned steps and a given (p, a) from $\mathcal{D}$. The goal is to find hard samples (p', a') so that the model has difficulty differentiating the pair (p, a') in the latent embedding space. We propose to select hard negatives by the similarity between p and p' to form a hard negative set. For a sample (p') from $\mathcal{D}$, we first calculate the similarity $sim(p, p')$ between p and p'. If $\alpha < sim(p, p') < \beta$, where α and β are hyperparameters, then p' will be added into the set $\mathcal{I}^-$. We don't want to select negative examples too close to the positive example, so we have $sim(p, p') < \beta$, and we don't want examples that are too easy, so we have $\alpha < sim(p, p')$. Based on manual observations, we set $\alpha = 0.3$ and $\beta = 0.7$. We use the cosine similarity function to measure the similarity of p and p'.

An illustration of how we construct the negative samples is shown on the left in Figure 2. Let $A(p)$ be the collection of all tail entities $\{a_j, a_j, ..., a_j\}$ from $\mathcal{D}$ such that each tail sequence has the same head p. For each $p_j \in \mathcal{I}^-$ we obtain the head sequence and tail sequence pairs $(p_j, a_{j,o})$, where $a_{j,o} \in A(p_j)$ is the collection of all tail entities $\{a_{j,1}, a_{j,2}, ..., a_{j,n}\}$ from $\mathcal{D}$ such that each tail sequence has the same head p_j. The union of these sets forms our hard negative set.

3.2 Positive Set Expansion

We propose to expand the positive set by utilizing the unique property of CSKGs to incorporate some potential while valuable positives.

Specifically, given a sequence pair of head and tail set (p, a_i), $a_i \in A(p)$, we measure the similarities of p with other head sequences p'. The p' with the highest similarity $sim(p, p')$ will be selected. Then, given the similar head sequence p', $A(p)$ and $A(p')$ may share some tail sequences. For example, in Figure 2, for the head sequences "PersonX loves cats, as a result, PersonX feels" and "PersonX likes cats, as a result, PersonX feels", both have same tail sequences while contain their own exclusive ones. Hence, we propose heuristically expanding the positive set A by inserting the missing tail sequences obtained from the tail sequence set $A(p')$.

3.3 Training Objective

For the sample (p_i, a_i), we use the InfoNCE loss with additive margin (Chen et al., 2020; Gao et al., 2021):

$$L_i = -\log \frac{e^{(\phi(\mathbf{p_i},\mathbf{a_h})-\gamma)/\tau}}{e^{(\phi(\mathbf{p_i},\mathbf{a_h})-\gamma)/\tau} + \sum_{\mathbf{j=1}}^{|\mathcal{I}^-|}\sum_{o=1}^{k} e^{\phi(\mathbf{p_i},\mathbf{a_{j,o}})/\tau}},$$

where the scoring function for a candidate sequence pair $\phi(\mathbf{p_i}, \mathbf{a_h}) = sim(\mathbf{p_i}, \mathbf{a_h})$. We use cosine similarity for our similarity function. For the hard positive, we select the one positive alternative a_h from the expanded set A which has the lowest similarity to p. The positive additive margin γ incentivizes the model to boost the score of the positive sequence pairs. By adjusting the temperature τ, the relative significance of negatives can be modified. A smaller value of τ increases the emphasis on challenging negatives, yet it also poses a risk of over-fitting to label noise.

3.4 Fine-Tuning Details

In practice, we fine-tune RoBERTa-Large (Liu et al., 2019) on the synthesized CSKG sequence pairs. The contrastive fine-tuning process directly equips the PLM with relation-aware commonsense knowledge, which can then be evaluated in zero-shot settings for commonsense QA and CSKG completion.

4 Experiments

In this section, we first introduce the CSKGs that we used in this study. Then we will present three evaluation tasks, unsupervised CSQA, inductive CSKG completion and claim verification, by intruding related benchmark datasets, baselines and main results. We conduct all experiments in a zero-shot setting, which means we do not have access to the official training data.

4.1 Commonsense Knowledge Graphs

Our experiments rely on two representative CSKGs, ConceptNet (Speer et al., 2017) and ATOMIC (Sap et al., 2019a). Each KG has different knowledge types. Following previous work(Wang et al., 2021; Su et al., 2022), we use CN-82K and ATOMIC in our experiments. The statistics are shown in Table 7. Details of the CSKGs are listed in Appendix A.1.

4.2 Unsupervised CSQA

In this section, we evaluate our framework on commonsense question answering datasets in an unsupervised way, which can be formalized as follows: given a question q and a set of answer candidates A, the model could choose the most likely candidate $\hat{a}$ by $\hat{a} = \arg\max_{a \in A} \text{sim}(\mathbf{q}, \mathbf{a})$, where $\mathbf{q}$ and $\mathbf{a}$ are representations obtained from the model.

Benchmarks: We conduct experiments on three different commonsense question answering datasets , COPA (Roemmele et al., 2011), SIQA (Sap et al., 2019b) and CSQA (Talmor et al., 2019) to verify the effectiveness of the proposed framework. Details of the datasets are listed in Appendix A.3.

Baselines: We compare the proposed framework with four different groups of baselines: (1) Vanilla PLMs (RoBERTa-Large (Liu et al., 2019), GPT2-L/M (Radford et al., 2019)); (2) Methods without relying on external CSKGs, instead by using PLMs to generate intermediate outputs (SEQA (Niu et al., 2021), self-talk (Shwartz et al., 2020), Dou (Dou and Peng, 2022)); (3) Prompting the large LMs to generate relevant knowledge given few-shot human annotations, including GKP (Liu et al., 2022) and TSGP (Sun et al., 2022); and (4) Models using CSKGs, including KTL (Banerjee and Baral, 2020), DynaGen (Bosselut et al., 2021), NLI-LM (Huang et al., 2021) and MICO (Su et al., 2022), a multi-view contrastive learning based baseline. For the details of each baseline method, please refer to their original papers. We are aware that there exist some other methods or method variants achieving better performance compared to the baselines listed here. However, they are either using larger backbone models (Sun et al., 2022) or trained with the larger even multiple knowledge bases (Ma et al., 2021; Kim et al., 2022). Both factors can improve the performance. Thus, we compare to methods with a similar model size as ours and the same knowledge bases. We also consider the issue of model size in §5.

Main Results: Table 1 shows the zero-shot evaluation results on benchmark datasets. Our model achieves the best performance across all baseline models on all datasets.

First, we compare our model with the vanilla PLMs, RoBERTa-Large (Liu et al., 2019), GPT2-L/M (Radford et al., 2019). It is not surprising that the LMs show significant and systematic performance gains on all datasets compared to the random baselines. Since it has been verified that the LMs already store implicitly vast amount of various types of knowledge in their parameters, such as relational and commonsense knowledge, which are universally indispensable for downstream tasks (Petroni et al., 2019).

Second, we compare our model with the methods generating intermediate outputs in the inference stage, such as SEQA (Niu et al., 2021) and self-talk (Shwartz et al., 2020). SEQA first generates a set of plausible answers and then compute the semantic similarity between each plausible answer and answer candidate. While self-talk iteratively queries the LMs with a set of information-seeking questions to disclose the potential background knowledge. However, this kind of methods cannot maintain their effectiveness systematically, even their performance is lower than the LM baselines. For example, as shown in Table 1, on CSQA dataset, self-talk is 8% lower than GPT2-Large, suggesting that self-talk may generate some spurious or misleading background knowledge. This shows that the explicit commonsense knowledge may be necessary to mitigate the hallucinations of LMs' generated knowledge. In light of this, our model injects explicit commonsense knowledge by self-supervising LMs on CSKGs. As shown in the results, our model can generate better commonsense knowledge representation advancing the unsupervised CSQA tasks.

Our method can achieve consistent improvement just by using relatively small backbone model. Compared with methods suzch as GKP (Liu et al., 2022) and TSGP (Sun et al., 2022), our best model outperforms them on SIQA and CSQA tasks without relying on large language models (LLMs). Similar as chain-of-thought (Wei et al., 2022), both GKP and TSGP first prompt the LLMs (GPT-3 and GPT2-XL, respectively) with few-shot human annotations to generate relevant background knowledge. However, knowledge snippets in nat-

Methods	Models	Knowledge Source	COPA dev	COPA test	SIQA dev	CSQA dev
Random	-	-	50.0	50.0	33.3	25.0
RoBERTa-L	RoBERTa-L	-	54.8	58.4	39.8	31.3
GPT2-L	GPT2-L	-	62.4	63.6	42.8	40.4
SEQA	GPT2-L	GPT2-L	-	-	46.6	34.6
self-talk	GPT2-[Distil/XL/L]	GPT2-[Distil/L/M]	66.0	-	46.2	32.4
Dou	ALBERT-XXL-v2	ALBERT-XXL-v2	-	-	44.1	50.9
GKP	T5-11b	few-shot exemplars + GPT-3	-	-	-	47.3
TSGP	GPT2-XL	few-shot exemplars + GPT2-XL	-	-	51.5	49.1
KTL	RoBERTa-L	ATOMIC	-	-	46.6	36.8
DynaGen	GPT2-M	COMET	-	-	50.1	-
NLI-LM	RoBERTa-L	ATOMIC+QNLI	-	-	-	52.1
MICO-CN	RoBERTa-L	ConceptNet	73.2	75.2	44.6	51.0
MICO-ATOMIC	RoBERTa-L	ATOMIC	79.4	77.4	56.0	44.2
Ours	RoBERTa-L	ConceptNet	73.8	77.2	46.2	**53.2**
Ours	RoBERTa-L	ATOMIC	**82.0**	**79.4**	**56.7**	47.8

Table 1: Accuracy (%) of unsupervised CSQA task on three public benchmarks. Our best scores are highlighted in bold.

ural language may not be sufficient to answer a commonsense-related question, since even LLMs still suffer from hallucination (Wei et al., 2022).

Our method can fine-tune LMs on CSKGs in a more effective and efficient way. Compared with methods using external CSKGs, such as KTL (Banerjee and Baral, 2020), DynaGen (Bosselut et al., 2021), NLI-LM (Huang et al., 2021) and MICO (Su et al., 2022), our method can achieve better performance even trained with the same CSKG. For a knowledge triplet, given knowledge representations of any two, KTL learns to generate the third one. While our method focuses on generating relation-aware contextualized representation given two sequence pairs. DynaGen dynamically generates contextually-relevant commonsense knowledge graphs by using a generative neural commonsense knowledge model, COMET (Bosselut et al., 2019). While the generated commonsense inferences are more context-relevant, it requires iterative generation that may impact the inference efficiency. Our method is more efficient by just generating contextually-relevant commonsense representations and selecting the most probable based on the largest similarity. NLI-LM utilizes extra NLI resources while unnecessary for our method. Our method outperform NLI-LM slightly by 1.1% on CSQA dataset. MICO is the most relevant to our method. It also utilizes contrastive multi-view training on CSKGs, while our method can bring consitent performance gains on all datasets compared with it. It shows the effectiveness of the two proposed modules, positive set expansion and hard negative sampling.

Model	ConceptNet MRR	ConceptNet Hits@10	ATOMIC MRR	ATOMIC Hits@10
ConvE	0.21	0.40	0.08	0.09
RotatE	0.32	0.50	0.10	0.12
Malaviya	12.29	19.36	0.02	0.07
InductivE	**18.15**	**29.37**	2.51	5.45
MICO	10.92	22.07	8.13	15.69
Ours	9.65	19.97	**8.29**	**15.93**

Table 2: Results on inductive CSKG completion. The best scores are highlighted in bold.

KG	Method	COPA dev	COPA test	SIQA dev	CSQA dev
Concept Net	Ours	73.8	77.2	**46.2**	**53.2**
	-w/o HNS	72.2	76.8	43.6	52.0
	-w/o PSE	**74.0**	**77.4**	43.9	52.7
ATOMIC	Ours	**82.0**	79.4	**56.7**	**47.8**
	-w/o HNS	79.0	**80.4**	56.0	44.4
	-w/o PSE	80.4	78.4	56.5	45.9

Table 3: Ablation study. The best scores are highlighted in bold.

4.3 Inductive CSKG Completion

Knowledge graphs, especially CSKGs, are often incomplete with missing entities and relations. Inductive CSKG completion evaluates the inductive capability of a model to predict relations triples for new, unseen entities (Wang et al., 2021). Given a knowledge triplet (h, r, t), the model needs to predict the unseen tail entity t by $(h, r, ?)$ or the unseen head entity by $(?, r^{-1}, t)$. Same as the previous work (Wang et al., 2021), we adopt the link predic-

Backbone	KG	COPA dev	COPA test	SIQA dev	CSQA dev
BERT Base	-	45.4	46.4	37.1	21.5
	ConceptNet	63.8	66.4	38.9	**43.2**
	ATOMIC	**69.8**	**74.0**	**48.2**	42.7
BERT Large	-	47.4	46.8	37.2	20.4
	ConceptNet	64.4	73.2	41.7	**47.8**
	ATOMIC	**73.2**	**74.2**	**51.6**	43.9
RoBERTa Base	-	52.0	55.2	38.4	29.2
	ConceptNet	62.4	69.6	40.1	**45.4**
	ATOMIC	**72.4**	**73.4**	**52.1**	41.0
RoBERTa Large	-	55.0	58.6	39.8	31.3
	ConceptNet	73.8	77.2	46.2	**53.2**
	ATOMIC	**82.0**	**79.4**	**56.7**	47.8

Table 4: Performance with different backbone LMs on unsupervised commonsense QA task.

tion task with standard evaluation metrics including MRR (Mean Reciprocal Rank) and Hits@10 to evalute the inductive CSKG completion models.

Benchmarks: In our experiments, following Wang et al. (2021), we use the inductive split of CN-82K and ATOMIC, where at least one of the entities in knowledge triplets of the testing sets is not present in the training set.

Baselines: We compare with ConvE (Dettmers et al., 2018), RotatE (Sun et al., 2019), Malaviya (Malaviya et al., 2020), InductivE (Wang et al., 2021) and MICO (Su et al., 2022).

Main Results: By training LMs with hard negative triplets and expanding the knowledge triplet with the potential missing alternatives on CSKGs, our method is able to generate superior commonsense knowledge representation, leading to the improved generalizability to unseen entities.

Table 2 shows the results of the inductive CSKG completion. Our method performs better on ATOMIC while remains comparable on ConceptNet. Previous entity embedding based methods by utilizing the existing entity links, such as ConvE (Dettmers et al., 2018) and RotatE (Sun et al., 2019), perform worse when it comes to the disconnected entities. For the graph neural network (GNN) based methods, such as Malaviya (Malaviya et al., 2020) and InductivE (Wang et al., 2021), by utilizing PLMs to initialize the entity embedding, the proposed GNNs trained on sampled subgraphs can significantly improve the generalizability on ConceptNet. However, the CSKGs are highly sparse and can be disconnected, the GNN-based methods could be failed when such a subgraph structure is not available (Franceschi et al., 2019).

In contrast, our method focuses on learning a relation-aware commonsense representation for each entity without relying on the graph structure. Same as MICO (Su et al., 2022), our method achieves better performance on ATOMIC while otherwise on ConceptNet compared with InductivE, one of the possible reasons could be the average length of the entity description in ATOMIC (6.12 words) is longer than that in ConceptNet (3.93 words). Longer sequences could enhance the PLMs to learn more accurate contextual representation for entity nodes. Compared with MICO, our method performs slightly worse on ConceptNet, one possible explanation is that more false negatives are introduced due to the hard negative sampling and positive set expansion.

5 Analysis

Ablation Study To further investigate what factors contribute to the performance gains, we conduct an ablation study by removing the step of hard negative sampling (HNS) and positive set expansion (PSE). Table 3 shows the results of ablation study on unsupervised CSQA task. Overall, when HNS or PSE is removed, the performance decreases on SIQA and CSQA whenever the model is trained with either ConceptNet or ATOMIC. Specifically, compared to the base model, training without HNS significantly hurts the performance by 2.6% and 0.7% on SIQA, which proves that hard negatives are effective in the existing contrastive learning instead of using in-batch negatives only. Meanwhile, removing PSE also degrades the performance most time, which shows that recovering the potential links between the head entity and the tail entity candidate by PSE contributes to learning superior commonsense-aware knowledge representation. However, removing PSE does not affect the accuracy much even can improve the performance slightly, which may be because that introducing PSE also incurs more false negatives in training.

Power of Scale We empirically test the influence of increasing the backbone LM size affecting the performance of the proposed model. Table 4 shows the results of different backbone LMs on unsupervised commonsense QA task. Overall, our method broadly benefits from backbone LM size increase. In addition, it conveys the same pattern as Table 1. ATOMIC benefits more for both COPA and SIQA, while ConceptNet is more helpful for CSQA.

6 Related Work

Contrastive Learning for NLP Contrastive learning has been applied into many NLP tasks. Such as, contrastive self-supervised objectives for text classification task (Fang et al., 2020; Kachuee et al., 2020); multi-view contrastive learning for dense encoder in open domain question answering (Karpukhin et al., 2020); sentence representation transfer with efficient contrastive framework (Yan et al., 2021; Gao et al., 2021). Among the works applying contrastive learning for NLP, Zhang and Stratos (2021) considered the importance of the hard negatives and proposed to combine hard negatives with appropriate score functions to improve the performance of zero-shot entity linking task. In this work, we propose to enhance contrastive learning with hard negative sampling for commonsense-aware knowledge representation task.

Unsupervised Commonsense Question Answering For the task of unsupervised CSQA, the vanilla PLMs can achieve moderate performance on most tasks. Furthermore, there are several methods generating intermediate outputs first by PLMs without relying on external CSKGs, such as SEQA (Niu et al., 2021), self-talk (Shwartz et al., 2020) and Dou (Dou and Peng, 2022). Some models incorporate CSKGs, including KTL (Banerjee and Baral, 2020), DynaGen (Bosselut et al., 2021), NLI-LM (Huang et al., 2021) and MICO (Su et al., 2022). Recently, a few methods prompt the large LMs to generate relevant knowledge given few-shot human annotations, including GKP (Liu et al., 2022) and TSGP (Sun et al., 2022). In this paper, we improve the commonsense knowledge representation by the sequence pairs synthesized CSKGs.

Commonsense Knowledge Graph Completion Existing KG completion methods can be adapted for CSKG completion, such as, ConvE (Dettmers et al., 2018) and RotatE (Sun et al., 2019) learn entity embeddings by the relation links between entity nodes. However, many entity nodes in CSKGs referring to the same concept are stored as distinct ones due to their free-form texts, resulting in larger and sparser graphs. To mitigate this issue, methods such as Malaviya (Malaviya et al., 2020) and InductivE (Wang et al., 2021), propose various graph neural network modules with the embeddings initialized from PLMs and focus on learn latent subgraph structures. Without leveraging graph structure, we also focus on the relation-aware knowledge representation with the free-form sequence pairs from CSKGS (Su et al., 2022).

7 Conclusion

In this paper, we propose to enhance the contrastive learning framework to fine-tune PLMs over CSKGs more effectively. Specifically, our method is divided into three steps: hard negative set sampling, positive set expansion and contrastive knowledge fine-tuning. We conduct extensive experiments on several unsupervised CSQA tasks and inductive CSKG completion with two widely used CSKGs, ConceptNet and ATOMIC. The performance gains demonstrate its effectiveness.

Limitations

First, in this paper, we focus on the commonsense knowledge representation learned on the synthesized sequence pairs from a given CSKG. However, the synthesized sequence pairs are missing contexts which may be indispensable for decision-making for some circumstances. Second, we propose to sample hard negatives during training instead of merely utilizing the in-batch negatives, which increases the memory footprint and computational costs. Third, we only focus on learning a relation-aware commonsense knowledge representation from the synthesized sequence pairs, while the subgraph structure of each entity node is also important for more fine-grained representation learning.

References

Pratyay Banerjee and Chitta Baral. 2020. Self-supervised knowledge triplet learning for zero-shot question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 151–162, Online. Association for Computational Linguistics.

Pratyay Banerjee, Swaroop Mishra, Kuntal Kumar Pal, Arindam Mitra, and Chitta Baral. 2021. Commonsense reasoning with implicit knowledge in natural language. In *3rd Conference on Automated Knowledge Base Construction*.

Antoine Bosselut, Ronan Le Bras, and Yejin Choi. 2021. Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*.

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, Florence, Italy. Association for Computational Linguistics.

Ruben Branco, António Branco, Joao Rodrigues, and Joao Silva. 2021. Shortcutted commonsense: Data spuriousness in deep learning of commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1521.

Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. 2016. Hard negative mining for metric learning based zero-shot classification. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 524–531. Springer.

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

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

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

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

Zi-Yi Dou and Nanyun Peng. 2022. Zero-shot commonsense question answering with cloze translation and consistency optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10572–10580.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.

Luca Franceschi, Mathias Niepert, Massimiliano Pontil, and Xiao He. 2019. Learning discrete structures for graph neural networks. In *International conference on machine learning*, pages 1972–1982. PMLR.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Canming Huang, Weinan He, and Yongmei Liu. 2021. Improving unsupervised commonsense reasoning using knowledge-enabled natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4875–4885, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Jinhao Ju, Deqing Yang, and Jingping Liu. 2022. Commonsense knowledge base completion with relational graph attention network and pre-trained language model. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4104–4108.

Mohammad Kachuee, Hao Yuan, Young-Bum Kim, and Sungjin Lee. 2020. Self-supervised contrastive learning for efficient user satisfaction prediction in conversational agents. *arXiv preprint arXiv:2010.11230*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Yu Jin Kim, Beong-woo Kwak, Youngwook Kim, Reinald Kim Amplayo, Seung-won Hwang, and Jinyoung Yeo. 2022. Modularized transfer learning with multiple knowledge graphs for zero-shot commonsense reasoning. *arXiv preprint arXiv:2206.03715*.

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, Berlin, Germany. Association for Computational Linguistics.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.

Yinhan Liu, 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*.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. 2021. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In *35th AAAI Conference on Artificial Intelligence*.

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.

Yilin Niu, Fei Huang, Jiaming Liang, Wenkai Chen, Xiaoyan Zhu, and Minlie Huang. 2021. A semantic-based method for unsupervised commonsense question answering. 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 3037–3049, Online. Association for Computational Linguistics.

Bo Ouyang, Wenbing Huang, Runfa Chen, Zhixing Tan, Yang Liu, Maosong Sun, and Jihong Zhu. 2021. Knowledge representation learning with contrastive completion coding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3061–3073, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Miao Peng, Ben Liu, Qianqian Xie, Wenjie Xu, Hua Wang, and Min Peng. 2022. Smile: Schema-augmented multi-level contrastive learning for knowledge graph link prediction. *arXiv preprint arXiv:2210.04870*.

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, Hong Kong, China. Association for Computational Linguistics.

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

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2020. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pages 90–95.

Maarten Sap, 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.

Maarten Sap, 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, Hong Kong, China. Association for Computational Linguistics.

Maarten Sap, 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*, pages 27–33.

Vered Shwartz, 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, Online. Association for Computational Linguistics.

Robert Speer and Catherine Havasi. 2013. Conceptnet 5: A large semantic network for relational knowledge. *The People's Web Meets NLP: Collaboratively Constructed Language Resources*, pages 161–176.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Ying Su, Zihao Wang, Tianqing Fang, Hongming Zhang, Yangqiu Song, and Tong Zhang. 2022. MICO: A multi-alternative contrastive learning framework for commonsense knowledge representation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1339–1351, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yueqing Sun, Yu Zhang, Le Qi, and Qi Shi. 2022. TSGP: Two-stage generative prompting for unsupervised commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 968–980, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Bin Wang, 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.

Liang Wang, Wei Zhao, Zhuoyu Wei, and Jingming Liu. 2022. Simkgc: Simple contrastive knowledge graph completion with pre-trained language models. *arXiv preprint arXiv:2203.02167*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. 2017. Sampling matters in deep embedding learning. In *Proceedings of the IEEE international conference on computer vision*, pages 2840–2848.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. *arXiv preprint arXiv:2105.11741*.

Lihui Zhang and Ruifan Li. 2022. Ke-gcl: Knowledge enhanced graph contrastive learning for commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 76–87.

Wenzheng Zhang and Karl Stratos. 2021. Understanding hard negatives in noise contrastive estimation. *arXiv preprint arXiv:2104.06245*.

A Details of CSKGs

A.1 CSKGs

Our experiments rely on two representative CSKGs, ConceptNet (Speer et al., 2017) and ATOMIC (Sap et al., 2019a).

ConceptNet. ConceptNet focuses on taxonomic, lexical and physical commonsense knowledge, describing the relation between a conceptual entity with another entity. Li et al. (2016) first introduced CN-100K which contains Open Mind Common Sense entries in the ConceptNet5 knowledge base (Speer and Havasi, 2013) to separate true and false triplets. However, the data split ratio of CN-100K is biased. In view of this issue, we use the new data split CN-82K proposed in (Wang et al., 2021) that is uniformly sampled.

ATOMIC. ATOMIC is an event-centric knowledge base, which contains everyday commonsense knowledge organized as nine typed *if-then* relations, e.g. xIntent, xWant. It focuses on different aspects of an event, such as social effect, mental states and causes. Following previous work, we use CN-82K and ATOMIC in our experiments (Wang et al., 2021; Su et al., 2022). The statistics are shown in Table 7.

A.2 Templates for Relation

Table 5 and Table 6 show the template for relation used for ATOMIC and ConceptNet, we adopted the version from InductivE [1].

A.3 Evaluation Benchmarks for Unsupervised CSQA

We evaluate our framework on commonsense question answering datasets, COPA (Roemmele et al., 2011), SIQA (Sap et al., 2019b) and CSQA (Talmor et al., 2019). We evaluate on both the dev and test splits unless the test split is hidden. The label information is only used for the final accuracy calculation.

COPA (Roemmele et al., 2011) COPA is a two-alternative commonsense causal reasoning dataset, where one alternative is more plausible than the other. We replace the term *cause* with *The cause for it was that* and *effect* with *As a result*, as in previous work (Su et al., 2022).[2]

[1]https://github.com/BinWang28/InductivE

[2]Please refer to Su et al. (2022) for more details.

Relation	rel template
xAttr	PersonX is seen as
xEffect	as a result, PersonX will
xWant	as a result, PersonX wants
xNeed	but before, PersonX needed
xReact	as a result, PersonX feels
xIntent	because PersonX wanted
oEffect	as a result, PersonY or others will
oReact	as a result, PersonY or others feel
oWant	as a result, PersonY or others want
xAttr rev	"PersonX is seen as", "because PersonX"
xEffect rev	"PersonX will", "because PersonX"
xWant rev	"PersonX wants", "because PersonX"
xNeed rev	"PersonX needs", "as a result PersonX"
xReact rev	"PersonX feels", "because PersonX"
xIntent rev	"PersonX wanted", "as a result PersonX"
oEffect rev	"PersonY or others will", "because PersonX"
oReact rev	"PersonY or others feel", "because PersonX"
oWant rev	"PersonY or others want", "because PersonX"

Table 5: Relation types and relation substitute templates from ATOMIC. *rev* mean reverse relation.

SIQA (Sap et al., 2019b) SIQA is three-choice dataset for testing social commonsense knowledge. Questions are built upon ATOMIC, focusing on social interactions about people's actions and their social implications.

CSQA (Talmor et al., 2019) CSQA is collected based on ConceptNet. Each question explores the potential taxonomic or physical commonsense relationships between entities and has five crowd-sourced candidate answers.

B Experimental Settings

We mainly run our experiments with RoBERTa-Large (Liu et al., 2019), which consists of 355M parameters. Our experiments are conducted with a A100 GPU. The running time of each experiment is about 5 10 hours. The results are averaged by three experiments.

Relation	relation templates
AtLocation	located or found at or in or on
CapableOf	is or are capable of
NotCapableOf	is not or are not capable of
Causes	causes
CausesDesire	makes someone want
CreatedBy	is created by
DefinedAs	is defined as
DesireOf	desires
Desires	desires
NotDesires	do not desire
HasA	has, possesses, or contains
HasFirstSubevent	begins with the event or action
HasLastSubevent	ends with the event or action
HasPrerequisite	to do this, one requires
HasProperty	can be characterized by being or having
InstanceOf	is an example or instance of
IsA	is a
MadeOf	is made of
MotivatedByGoal	is a step towards accomplishing the goal
PartOf	is a part of
ReceivesAction	can receive or be affected by the action
SymbolOf	is a symbol of
UsedFor	used for
LocatedNear	is located near
RelatedTo	is related to
InheritsFrom	inherits from
LocationOfAction	is acted at the location of
HasPainIntensity	causes pain intensity of
AtLocation rev	is the position of
CapableOf rev	is a skill of
NotCapableOf rev	is not a skill of
Causes rev	because
CausesDesire rev	because
CreatedBy rev	create
DefinedAs rev	is known as
DesireOf rev	is desired by
Desires rev	is desired by
NotDesires rev	is not desired by
HasA rev	is possessed by
HasFirstSubevent rev	is the beginning of
HasLastSubevent rev	is the end of
HasPrerequisite rev	is the prerequisite of
HasProperty rev	is the property of
InstanceOf rev	include
IsA inversed	includes
MadeOf rev	make up of
MotivatedByGoal rev	motivate
PartOf rev	include
ReceivesAction rev	affect
SymbolOf rev	can be represented by
UsedFor rev	could make use of
LocatedNear rev	is located near
RelatedTo inversed	is related to
InheritsFrom rev	hands down to
LocationOfAction rev	is the location for acting
HasPainIntensity rev	is the pain intensity caused by

Table 6: Relation types and relation substitute templates from ConceptNet. *rev* mean reverse relation.

Dataset	Entities	Relations	Train Edges	Valid Edges	Test Edges	Avg. In-Degree
ConceptNet	78,334	34	81,920	9,795	9,796	1.31
ATOMIC	304,388	9	610,536	24,355	24,486	2.58

Table 7: Distribution of train, valid, and test edges from CN-82K and ATOMIC. Avg. In-Degree is the average number of tail entity connected to head entity.