MICO: A Multi-alternative Contrastive Learning Framework for Commonsense Knowledge Representation

Ying Su\textsuperscript{1}, Zihao Wang\textsuperscript{1}, Tianqing Fang\textsuperscript{1}, Hongming Zhang\textsuperscript{2}, Yangqiu Song\textsuperscript{1}, Tong Zhang\textsuperscript{1}

\textsuperscript{1}HKUST, \textsuperscript{2}Tencent AI lab, Seattle

\{ysuay,zwanggc\}@connect.ust.hk, \{tfangaa,yqsong\}@cse.ust.hk, hongmzhang@global.tencent.com, tongzhang@ust.hk

Abstract

Commonsense reasoning tasks such as commonsense knowledge graph completion and commonsense question answering require powerful representation learning. In this paper, we propose to learn commonsense knowledge representation by MICO, a Multi-alternative contrastive learning framework on Commonsense knowledge graphs (MICO). MICO generates the commonsense knowledge representation by contextual interaction between entity nodes and relations with multi-alternative contrastive learning. In MICO, the head and tail entities in an \((h,r,t)\) knowledge triple are converted to two relation-aware sequence pairs (a premise and an alternative) in the form of natural language. Semantic representations generated by MICO can benefit the following two tasks by simply comparing the distance score between the representations: 1) zero-shot commonsense question answering task; 2) inductive commonsense knowledge graph completion task. Extensive experiments show the effectiveness of our method.

1 Introduction

Commonsense reasoning is a fundamental problem in artificial intelligence. Recently in the NLP field, much attention has been paid to commonsense reasoning in the following two aspects. First, more commonsense knowledge graphs (CKGs) (Sap et al., 2019a; Fang et al., 2021) were developed to support new types of reasoning tasks, such as commonsense knowledge graph completion (CKGC) (Malaviya et al., 2020). Another way to evaluate machine learning models’ commonsense reasoning capabilities is using commonsense question answering (CQA) tasks (Zellers et al., 2018; Sap et al., 2019b; Bisk et al., 2020). Existing approaches to deal with the above problems commonly involve fine-tuning large pre-trained language models, such as BERT (Kenton and Toutanova, 2019), RoBERTa (Liu et al., 2019), and GPT2 (Radford et al., 2019), by either incorporating the entire knowledge base for CKGC (Yao et al., 2019; Bosselut et al., 2019) or injecting the knowledge base to provide background knowledge for zero-shot CQA (Banerjee and Baral, 2020; Bosselut et al., 2021; Ma et al., 2021).

In fact, both CKGC and zero-shot CQA can be formulated in a unified way, where a question can be constructed based on the head entity and relation in a knowledge graph, and then finding the tail entity, which is regarded as an answer, based on the constructed question. In this way, incorporating the entire knowledge base for CKGC and injecting the KG in pre-trained LMs for zero-shot CQA can be unified as a semantic matching problem, where a powerful representation learning for the matching becomes the most important problem. This also means that, after we unify them for CKGC and CQA in the same way, we can perform zero-shot CQA by simply leveraging the model finetuned on the entire CKGs for CKGC.

Existing commonsense-related representation learning usually leverage a CKG embedding framework (Malaviya et al., 2020; Wang et al., 2021), or fine-tuning a generative language model (Bosselut et al., 2019). However, they were not aware of the challenges that a typical CKG brings. First, in a typical CKG, such as ConceptNet (Liu and Singh, 2004) and ATOMIC (Sap et al., 2019a), nodes are loosely structured free-from texts, which means that previous embedding based on negative sampling cannot substantially support sufficient training because of sparsity. On the other hand, a generative model can only take positive examples for training so the capability of determining the negative answers is limited.

In this paper, we propose a new framework called MICO, a Multi-alternative contrastive learning framework for Commonsense knowledge representation. The representations can benefit across tasks by easily calculating the semantic distances
with unified vector representations. In this way, though many distinct nodes may have similar concepts (Wang et al., 2021), they are still close in the semantic space. Figure 1 shows an example of this advantage on CQA. The entity node PersonX ask if PersonY had seen related to the right answer is not directly connected to the nodes PersonX leaves PersonX’s book or PersonX leaves PersonY’s book but these nodes share similar tail entity nodes. Therefore, the right answer can be found by semantic matching as it is close to the given context and question in the semantic space.

To unify the form of CKGC and CQA, we follow the idea in COPA (Roemmele et al., 2011) where commonsense causal reasoning can be evaluated by selecting the most plausible alternatives given the premise. We first converts the knowledge triplets \((h, r, t)\) into sequence pairs \((P, A)\) \((P\) for premise and \(A\) for alternative). MICO then encodes the sequence pairs into embeddings and measures their distance by a similarity function as we assume the representations of related knowledge lie close in the embedding space. Furthermore, we enhance the representation learning by a contrastive loss with sufficient sampling over the sparse CKG. The alternative from the same triplet is a positive sample to the premise under the contrastive learning framework. Alternatives from other knowledge triplets with different premises are negative samples. MICO also takes consideration of the structure from CKGs, where one head node \(h\) may connect to several tail nodes \(t\) under the same relation \(r\). MICO dynamically selects a hard alternative from multi-alternatives for a premise during training.

Experiments on two typical commonsense knowledge graphs and two types of tasks, zero-shot CQA and inductive CKGC, demonstrate the effectiveness of our methodology. Our code is open-sourced.\(^1\)

## 2 Related Work

### 2.1 Commonsense Question Answering

Background knowledge is necessary for commonsense question answering tasks. Many researches resort to knowledge bases for background knowledge. The works towards this direction can be mainly classified into two streams: incorporating the knowledge base for zero-shot CQA (Yang et al., 2019; Banerjee and Baral, 2020; Bosselut et al., 2021; Ma et al., 2021) or retrieving the related knowledge from the knowledge base for task-specific CQA (Paul and Frank, 2019; Lin et al., 2019; Feng et al., 2020; Lv et al., 2020; Yasunaga et al., 2021; Xu et al., 2021; Zhang et al., 2021).

Among the works in incorporating the knowledge base for zero-shot CQA, COMET-DynaGen (Bosselut et al., 2021) aggregates all paths of generated commonsense knowledge to the answers from commonsense transformer COMET (Bosselut et al., 2019) trained on CKGs. KTL (Banerjee and Baral, 2020) encodes the knowledge triplets from CKGs into pre-trained LMs by learning triplet representation, aiming to complete a knowledge triplet given the other two. Unlike KTL, we target enhancing the relation-aware representation learning in the form of natural language sequence pairs.

### 2.2 CKG Knowledge Representation

Knowledge representation from knowledge graphs (KGs) has significantly progressed and benefited the KG completion task. Typical methods for KG completion tasks are mainly embedding-based, which utilize the structural information observed in the knowledge triplets (Nickel et al., 2011; Bordes et al., 2013; Wang et al., 2014; Trouillon et al., 2016; Toutanova et al., 2015; Sun et al., 2019). Recent researches also show that external information, such as the textual description of nodes or relation descriptions, can help boost the performance on the task, like ConvE (Dettmers et al., 2018) and ConvTransE (Shang et al., 2019). To transfer the knowledge from pre-trained LMs into knowledge

---

\(^1\)https://github.com/HKUST-KnowComp/MICO

---

![Figure 1: An example from a CQA task (SIQA (Sap et al., 2019b)) and related knowledge in a CKG (ATOMIC).](image-url)
Figure 2: Overview framework of MICO. The knowledge triplets are first sampled and converted into sequence pairs. A transformer block encodes the sequence pairs into knowledge embeddings. The contrastive loss updates the parameters of the transformer based on the contrast between the selected positive and negative tail sequences given a head sequence.

Figure 3: An example of converting knowledge triplet \((h, r, t)\) to sequence pairs \((P, A)\) from ATOMIC. \(r^{-1}\) is the reverse relation of \(r\).

3.1 Knowledge Triplets to Sequence Pairs

A commonsense knowledge graph is denoted as \(G = \{V, E, R\}\), where \(V\) is the set of entities, \(E\) is the set of edges, and \(R\) is the set of relations. Knowledge triplet \(e \in E\) is composed by \((h, r, t)\) where head entity \(h\) and tail entity \(t\) are entities from \(V\) and \(r\) is from \(R\). Each entity comes with a free-form text description.

To convert the knowledge triplet into sequence pairs as inputs to MICO, we substitute the relation with human-readable language templates and connect it to entities. Typically relations are represented as specific words or short phrases in the CKG, for example, \(x\text{Want}\) from ATOMIC and \(At\text{Location}\) from ConceptNet. Following Hwang et al. (2021), we design natural language templates to replace the original relations and connect them to entities, forming context-aware sequences. An example from ATOMIC is shown in Figure 3. The template for \(x\text{Want}\) would be \(\text{as a result, PersonX wants}\). We also design a template for its reverse version \(r^{-1}\) so it can be connected to the tail entity to form an additional sequence pair. Details of substitute templates are listed in Appendix A.1. We denote the newly constructed sequence pair as \((P, A)\). For a premise \(P\), there may be multiple alternatives connected to it, denoted as \(\{A_1, A_2, A_3, ...\}\).

3.2 MICO

MICO is a multi-alternative contrastive learning framework for commonsense knowledge representation with knowledge sequence pairs as inputs. Recent researches have greatly progressed in sequence representation learning by contrastive learning (Carlsson et al., 2020; He et al., 2020; Gao et al., 2021).
2021), in which positive sequence pairs are considered semantic related and are close neighbors in the embedding space. MICO follows the idea and minimizes the distance between the premise $P$ and its connected alternative $A$.

First, a transformer encodes the constructed sequence pairs $(P, A)$ into embeddings to extract initial representations. Specifically, when using BERT as the transformer, BERT-specific start and end tokens are padded to the input sequence: $x = x_0, x_1, \ldots, x_n$ is converted to $x = [CLS], x_0, x_1, \ldots, x_n, [SEP]$. The sequence pairs are then transformed to token pairs as $P_{tok}$ and $A_{tok}$ by a transformer tokenizer. To get the initial representation, a transformer encoder encodes the token pairs as:

$$E_p = \text{Encoder}(P_{tok}), \quad E_a = \text{Encoder}(A_{tok}),$$  \hspace{1cm} (1)

where $E_h$ and $E_t$ are hidden states of the last layer. The representation of the hidden state for the $[CLS]$ token is used as the representation of the input sequence. For the head and tail sequences, representations are:

$$s = E_p[0], \quad g = E_a[0].$$  \hspace{1cm} (2)

As we assume the sequence pairs lie close in the embedding space, we use a similarity function to measure the distance between sequence pairs. The function $f$ can be cosine similarity or dot product:

$$sim(s, g) = f(s, g).$$  \hspace{1cm} (3)

For a premise $P$, its paired alternative $A$ is a positive sample while alternatives paired with other premises are negative samples in the same batch during training. To minimize the semantic distance between the $i$-th sequence pair in a batch, the contrastive loss is:

$$\ell_i = -\log \frac{e^{sim(s_i, g_i^+)} / \tau}{\sum_{j=1}^{N} e^{sim(s_i, g_j^+)} / \tau},$$  \hspace{1cm} (4)

where $N$ is the batch size and $\tau$ is the temperature parameter.

Many research efforts aim to improve representation learning by generating multiple views for the same sample as data augmentation in multi-view contrastive learning approaches (Bachman et al., 2019; Tian et al., 2020; Niu et al., 2022). In CKG, a sequence head $P$ may have multiple positive tails $\{A_1, A_2, A_3, \ldots\}$. Inspired by multi-view contrastive learning, we propose a multi-alternative framework to utilize the multiple positive alternatives for improving the learning of commonsense knowledge representation. Specifically, we dynamically sample a hard positive alternative from multiple alternatives during training.

The representations generated for the premise with its alternatives are $s$ and $\{g_1^+, g_2^+, g_3^+, \ldots\}$. Among the multiple positive alternatives, the one with the largest distance to the premise is selected as the hard positive. Because we aim to minimize the semantic similarity between the premise and alternatives, selecting the alternative with the least similarity would increase the training loss:

$$g_p = \min\{sim(s, g_1^+), \ldots, sim(s, g_k^+)) \},$$  \hspace{1cm} (5)

where $k$ is the number of candidate alternatives during training. The new contrastive loss for $i$-th sequence pairs during training is:

$$L_i = -\log \frac{e^{sim(s_i, g_p^+)} / \tau}{\sum_{j=1}^{k} e^{sim(s_i, g_j^+)} / \tau} + \delta_{ij} \sum_{j=1}^{k} e^{sim(s_i, g_j^+)} / \tau,$$  \hspace{1cm} (6)

where $\delta_{ij} \in \{0, 1\}$ is an indicator that equals 1 if $i \neq j$, and $g_{j,o}^+$ is the $o$-th positive tail of $j$-th sample in the batch.

### 4 Experiments

In this section, we first introduce the CKGs used as knowledge sources, and then two kinds of evaluation tasks (zero-shot CQA and inductive CKGC). Finally, we introduce baseline methods for the two tasks separately.

#### 4.1 CKG

We conduct experiments on two typical commonsense knowledge graphs, ConceptNet (Speer et al., 2017) and ATOMIC (Sap et al., 2019a).

**ConceptNet.** ConceptNet has been the most fundamental commonsense knowledge graph over the past decade (Liu and Singh, 2004). CN-100K was built on the knowledge triplets in ConceptNet and first introduced in Li et al. (2016). It contains Open Mind Common Sense (OMCS) (Singh et al., 2002) entries in the ConceptNet5 (Speer et al., 2017). CN-82K is a uniformly sampled version of CN-100K dataset which contains more unseen entities in the test split (Wang et al., 2021).
### Table 1: Distribution of train, valid, and test sequence pairs from ConceptNet and ATOMIC.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Entities</th>
<th>Relations</th>
<th>Train Pair</th>
<th>Valid Pair</th>
<th>Test Pair</th>
<th>Avg Degree</th>
<th>Avg Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConceptNet</td>
<td>78,334</td>
<td>34</td>
<td>163,840</td>
<td>19,590</td>
<td>19,592</td>
<td>1.87</td>
<td>3.93</td>
</tr>
<tr>
<td>ATOMIC</td>
<td>304,388</td>
<td>9</td>
<td>1,221,072</td>
<td>48,710</td>
<td>48,972</td>
<td>2.52</td>
<td>6.12</td>
</tr>
</tbody>
</table>

*Avg Degree* is the average number of tail sequence connected to head sequence and *Avg Words* is the average words number for head sequence and tail sequence.

**ATOMIC.** ATOMIC (Sap et al., 2019a) contains rich social commonsense knowledge about day-to-day events. The dataset specifies the effects, needs, intents, and attributes of the actors in the events, covering nine relations and 877k knowledge tuples. Dataset built from ATOMIC for CSKG completion is first created in Malaviya et al. (2020).

In our experiments, we follow Wang et al. (2021) to use CN-82K and ATOMIC. To better evaluate the generalizability of representation from MICO, we conduct experiments with the inductive splits in which one of entity nodes in a knowledge triplet from the valid and test split does not appear in the training dataset. Statistics of the converted sequence pairs from original datasets are shown in Table 1.

### 4.2 Evaluation Tasks

Based on that CKGC and CQA can be unified into the same form of selecting alternatives given a premise, commonsense knowledge representations generated from MICO are evaluated on these two tasks.

#### 4.2.1 Zero-shot CQA

The knowledge representation is evaluated on three multiple-choice CQA tasks: COPA (Roemmele et al., 2011), SIQA (Sap et al., 2019b), and CSQA (Talmor et al., 2019). Accuracy is used as the evaluation metric. For each task, the query composed by context and question can be converted into the form as a premise. The answers are viewed as possible plausible alternatives. In this way, the multiple-choice question can be solved by selecting the closet representation pairs generated from MICO given the query and candidate answers. We denote the representation for query as \( q \) and candidate answers as \( \{a_1, a_2, \ldots, a_m\} \). The answer with the highest score is the predicted answer \( i^* \), where

\[
    i^* = \arg \max_{i=1,...,m} \text{sim}(q, a_i). \tag{7}
\]

**COPA.** Choice of Plausible Alternatives is a two-way multiple-choice commonsense reasoning task between events. COPA consists of 1,000 questions, 500 for the development set and 500 for the test set. To make the form of relation consistent with the training dataset in natural language, we substitute cause as *The cause for it was that* and effect as *As a result.*

**SIQA.** The queries in Social IQA are collected based on ATOMIC. Each question in the dataset describes social interactions and has three crowdsourced candidate answers. The dataset’s development split and test split are used as zero-shot evaluation, containing 1,954 and 2,059 questions, respectively.

**CSQA.** The questions in CommonsenseQA are general questions about concepts in ConceptNet. Each question has five candidate answers. The development set is used as evaluation set, containing 1,221 questions.

#### 4.2.2 Inductive CKGC

Inductive CKGC is an important task for CKG because unseen entity nodes are introduced in real-world CKGC from time to time and many distinct nodes may refer to same concept due to their free-form text description (Wang et al., 2021). In the inductive CKGC task, at least one of the nodes in knowledge triplets is not shown in the training dataset. Following Wang et al. (2021), each triplet \((h, r, t)\) is measured in two directions: \((h, r, \cdot)\) and \((t, r^{-1}, \cdot)\). Inverse relations \(r^{-1}\) are added as additional relation types. We use the MRR (mean reciprocal rank) and Hits@10 score as the evaluation metrics.

### 4.3 Baselines

For zero-shot CQA tasks, we compare our method with baselines including pre-trained LMs (RoBERTa (Liu et al., 2019), GPT2 (Radford et al., 2019)), using pre-trained LMs as knowledge sources (self-talk (Shwartz et al., 2020), Dou (Dou and Peng, 2022)), and pre-trained LMs trained on CKGs (KTL (Banerjee and Baral, 2020), COMET-
Table 2: Results on Zero-shot CQA tasks. COMET is the commonsense transformer trained on ATOMIC. For MICO, k is set as 2. RoBERTa-L and GPT2-M have comparable parameter size. KRL is the knowledge representation method in KTL.

DynaGen (Bosselut et al., 2021)).

For inductive CKGC tasks, we compare our method with ConvE (Dettmers et al., 2018), Ro-
atatE (Sun et al., 2018), Malaviya (Malaviya et al., 2020) and InductivE (Wang et al., 2021). More
details about baseline models for the two tasks are introduced in Appendix A.2.

4.4 Implementation Details

Our experiments are run on RTX A6000. Each
experiment is run on a single GPU card. The training batch size is 196. Max sequence length for training is 32. The learning rate is set as 1e-5 for Bert-base and RoBERTa-base. For RoBERTa-large (RoBERTa-L), the learning rate is set as 5e-6. We use AdamW (Loshchilov and Hutter, 2018) optimizer. For experiments with the MICO framework, \(\tau\) is set as 0.07. The valid set is evaluated by con- strastive loss metric and used to select a best model for further evaluation. The models are trained for 10 epochs and early stopped when the change of validation loss is within 1%.

5 Results

5.1 Main Results

The main results include MICO on the zero-shot CQA and the inductive CKGC.

5.1.1 Results on Zero-shot CQA

The results on CQA tasks are shown in Table 2. Baseline systems based on pre-trained language models such as RoBERTa-L and GPT2-L provides strong baselines. Simply comparing the language model score from RoBERTa-L or GPT2-L outperforms random guess by a large margin. This shows that pre-trained language models are encoded with useful knowledge which can benefit the CQA tasks.

MICO generates knowledge representation encoded with commonsense knowledge by the fine-tuned LMs with self-supervision signal from CKGs. Compared with methods such as self-talk (Shwartz et al., 2020) and Dou (Dou and Peng, 2022), our method outperforms all the evaluation datasets. Self-talk and Dou utilize the pre-trained language models as knowledge source and mine relevant knowledge that may benefit the CQA tasks. However, such knowledge is still not sufficient. By fine-tuning on CKG, MICO can successfully inject the commonsense knowledge into pre-trained LMs and generate meaningful representation benefiting the CQA tasks.

Another finding is that the representation generated from MICO can be easily generalized to out-of-domain datasets. SIQA achieves best results when ATOMIC used as the knowledge source and
Table 3: Results on inductive CKGC. MICO∗ for BERT-base, MICO♢ for RoBERTa-base, MICO♡ for RoBERTa-L. k is set as 2.

Table 4: Backbone model study on two CKGs and evaluation on CQA tasks. For MICO, k is set as 2 during training.
mostly increase as $k$ increases. This indicates that larger $k$ helps the model generalize better in pairing the in-domain knowledge sequences. However for ConceptNet, the performance drops when $k$ is greater than 3. The limited average degree of nodes in ConceptNet may explain this as larger $k$ does not induce new candidate tails. Therefore, the model tends to fit the seen nodes better.

### 5.2.3 Sampling Strategy
We analyze the influence of selecting a hard positive compared with randomly sampling a positive from candidate sets. The results are show in Table 7. The experiments are conducted on the backbone model with BERT-base and $k = 2$. Compared to random sampling, MICO mostly outperforms on the three datasets. This indicates that the hard positive during training can benefit the generalization of the representation. The only exception is training on ATOMIC and testing on SIQA. The possible explanation is that there is possibly a distribution gap between the training dataset and SIQA dataset. However, generally our sampling strategy can improve the generalizability of the representation.

### 6 Discussion
This section shows that transformers from MICO can construct commonsense representations for CKG and benefit commonsense knowledge retrieval given queries. One example from SIQA and retrieved possible alternatives from ATOMIC is shown in Table 6. We first encode all the alternatives in CKG by the transformer finetuned with ATOMIC and original transformer without any fine-tuning. The top 5 retrieved nodes are listed by the ranks of similarity score in descending order. We can find that the transformer finetuned on CKG can successfully pair the query with reasonable alternatives from CKG compared to original pre-trained transformer. Therefore, our method provides an efficient way to collect the related knowledge from CKG and may benefit the researches which require retrieved implicit background knowledge to reason over.

However, the representation generated from MICO still has some drawbacks as shown in the results. “Jordan look for it at library” would be a reasonable node instead of “Jordan look for it at home”. This shows that the representation still need future work to distinguish the detailed concepts.

### 7 Conclusion
In this paper, we propose a MICO, a multi-alternative contrastive learning framework over commonsense knowledge graphs to learn com-
monsense knowledge representation. The framework converts the knowledge triplets into sequence pairs and learns superior knowledge representation through contrastive learning techniques. The generated representations perform well over zero-shot CQA tasks and inductive CKGC tasks. Furthermore, for CQA tasks, the related knowledge can be provided by simply retrieving the commonsense knowledge representations of CKGs.

Acknowledgements

The authors of this paper were supported by the NSFC of China, the RIF (R6020-19 and R6021-20) and the GRF (16211520) from RGC of Hong Kong, the MHKJS (MHP/001/19) from ITC of Hong Kong and the National Key R&D Program of China (2019YFE0198200) with special thanks to HKMAAC and CUSBLT, and the Jiangsu Province Science and Technology Collaboration Fund (BZ2021065). We also thank the support from the UGC Research Matching Grants (RMGS20EG01-D, RMGS20CR11, RMGS20CR12, RMGS20EG19, RMGS20EG21).

Limitations

As shown in the discussion, the commonsense knowledge representation generated from MICO can capture the rough meanings of the whole word sequences. While for detailed concepts, the representation failed to distinguish since the representation is extracted from a specific token to represent the meaning of the whole word sequence. However, concepts are key elements in semantics so future work is still needed to improve the representation.

References


Table 9: Relation types and relation substitute templates from ATOMIC. rev mean reverse relation.

<table>
<thead>
<tr>
<th>Relation</th>
<th>rel template</th>
</tr>
</thead>
<tbody>
<tr>
<td>xAttr</td>
<td>PersonX is seen as</td>
</tr>
<tr>
<td>xEffect</td>
<td>as a result, PersonX will</td>
</tr>
<tr>
<td>xWant</td>
<td>as a result, PersonX wants</td>
</tr>
<tr>
<td>xNeed</td>
<td>but before, PersonX needed</td>
</tr>
<tr>
<td>xReact</td>
<td>as a result, PersonX feels</td>
</tr>
<tr>
<td>xIntent</td>
<td>because PersonX wanted</td>
</tr>
<tr>
<td>oEffect</td>
<td>as a result, PersonY or others will</td>
</tr>
<tr>
<td>oReact</td>
<td>as a result, PersonY or others feel</td>
</tr>
<tr>
<td>oWant</td>
<td>as a result, PersonY or others want</td>
</tr>
</tbody>
</table>

Table 9: Relation types and relation substitute templates from ATOMIC. rev mean reverse relation.

A Appendix

A.1 Templates for Relation

For training dataset of two CSKGs, we used the version from InductivE

A.1.1 ATOMIC

In ATOMIC, there are nine relations. The substitute template of original relations and reverse relations are shown in Table 9. PersonX and PersonY are substituted by "John" or "Tom" respectively.

A.1.2 ConceptNet

ConceptNet contains 34 relations. The substitute template of original relations and reverse relations are shown in Table 10.

A.2 Baselines

A.2.1 Commonsense Question Answering

Self-Talk (Shwartz et al., 2020). Self-Talk inquires LMs for implicit background knowledge to solve multiple-choice commonsense tasks. The model uses pretrained LMs as knowledge sources.

Dou (Dou and Peng, 2022). Dou extracts the related background knowledge embedded in pretrained LMs by designing fill-in-the-blank prompts for commonsense question answering tasks. We compare with the syntactic-based rewriting method in which no supervision from curated-annotated training data is used.

KTL (Banerjee and Baral, 2020). Two ways are used to learn from knowledge triplets from knowledge graphs which can be used to perform zero-shot QA, namely knowledge representation learning (KRL) and span masked language modeling (SMLM). We compare with the KRL method.

COMET-DynaGen (Bosselut et al., 2021). COMET-DG implements zero-shot commonsense QA by inference over dynamically generated commonsense knowledge graphs as related knowledge from COMET.

A.2.2 Inductive CSKG Completion

ConvE (Dettmers et al., 2018). ConvE stacks the node embedding and relation embedding and reshapes the resulting tensor into the same dimensionality as the node embeddings by a 2D convolution operation.

RotatE (Sun et al., 2018). RotatE utilizes the rotation operation and defines the distance function between entities and relation as \(d_r(h, t) = \|h \circ r - t\|\).

Malaviya (Malaviya et al., 2020). The method adopts a graph convolutional neural network GCN to learn graph structure information and a pretrained LMs to represent contextual knowledge.

InductivE (Wang et al., 2021). The model directly computes entity embeddings from raw attributes and a GCN decoder with a novel densification process to enhance unseen entity representation with neighboring structural information.

---

<table>
<thead>
<tr>
<th>Relation</th>
<th>relation templates</th>
</tr>
</thead>
<tbody>
<tr>
<td>AtLocation</td>
<td>located or found at or in or on</td>
</tr>
<tr>
<td>CapableOf</td>
<td>is or are capable of</td>
</tr>
<tr>
<td>NotCapableOf</td>
<td>is not or are not capable of</td>
</tr>
<tr>
<td>Causes</td>
<td>causes</td>
</tr>
<tr>
<td>CausesDesire</td>
<td>makes someone want</td>
</tr>
<tr>
<td>CreatedBy</td>
<td>is created by</td>
</tr>
<tr>
<td>DefinedAs</td>
<td>is defined as</td>
</tr>
<tr>
<td>DesireOf</td>
<td>desires</td>
</tr>
<tr>
<td>Desires</td>
<td>desires</td>
</tr>
<tr>
<td>NotDesires</td>
<td>do not desire</td>
</tr>
<tr>
<td>HasA</td>
<td>has, possesses, or contains</td>
</tr>
<tr>
<td>HasFirstSubevent</td>
<td>begins with the event or action</td>
</tr>
<tr>
<td>HasLastSubevent</td>
<td>ends with the event or action</td>
</tr>
<tr>
<td>HasPrerequisite</td>
<td>to do this, one requires</td>
</tr>
<tr>
<td>HasProperty</td>
<td>can be characterized by being or having</td>
</tr>
<tr>
<td>InstanceOf</td>
<td>is an example or instance of</td>
</tr>
<tr>
<td>IsA</td>
<td>is a</td>
</tr>
<tr>
<td>MadeOf</td>
<td>is made of</td>
</tr>
<tr>
<td>MotivatedByGoal</td>
<td>is a step towards accomplishing the goal</td>
</tr>
<tr>
<td>PartOf</td>
<td>is a part of</td>
</tr>
<tr>
<td>ReceivesAction</td>
<td>can receive or be affected by the action</td>
</tr>
<tr>
<td>SymbolOf</td>
<td>is a symbol of</td>
</tr>
<tr>
<td>UsedFor</td>
<td>used for</td>
</tr>
<tr>
<td>LocatedNear</td>
<td>is located near</td>
</tr>
<tr>
<td>RelatedTo</td>
<td>is related to</td>
</tr>
<tr>
<td>InheritsFrom</td>
<td>inherits from</td>
</tr>
<tr>
<td>LocationOfAction</td>
<td>is acted at the location of</td>
</tr>
<tr>
<td>HasPainIntensity</td>
<td>causes pain intensity of</td>
</tr>
</tbody>
</table>

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 10: Relation types and relation substitute templates from ConceptNet. rev mean reverse relation.