Dense-ATOMIC: Towards Densely-connected ATOMIC with High Knowledge Coverage and Massive Multi-hop Paths

Xiangqing Shen, Siwei Wu, and Rui Xia*
School of Computer Science and Engineering,
Nanjing University of Science and Technology, China
{xiangqing.shen, wusiwei, rxia}@njust.edu.cn

Abstract

ATOMIC is a large-scale commonsense knowledge graph (CSKG) containing everyday *if-then* knowledge triplets, i.e., {*head event*, relation, *tail event*}. The one-hop annotation manner made ATOMIC a set of independent bipartite graphs, which ignored the numerous links between events in different bipartite graphs and consequently caused shortages in knowledge coverage and multi-hop paths. In this work, we aim to construct Dense-ATOMIC with high knowledge coverage and massive multi-hop paths. The events in ATOMIC are normalized to a consistent pattern at first. We then propose a CSKG completion method called Rel-CSKGC to predict the relation given the *head event* and the *tail event* of a triplet, and train a CSKG completion model based on existing triplets in ATOMIC. We finally utilize the model to complete the missing links in ATOMIC and accordingly construct Dense-ATOMIC. Both automatic and human evaluation on an annotated subgraph of ATOMIC demonstrate the advantage of Rel-CSKGC over strong baselines. We further conduct extensive evaluations on Dense-ATOMIC in terms of statistics, human evaluation, and simple downstream tasks, all proving Dense-ATOMIC's advantages in Knowledge Coverage and Multi-hop Paths. Both the source code of Rel-CSKGC and Dense-ATOMIC are publicly available on https://github.com/NUSTM/Dense-ATOMIC.

1 Introduction

ATOMIC is a large-scale human-annotated commonsense knowledge graph focusing on the inferential knowledge in social life (Sap et al., 2019). It consists of nine *if-then* relation types describing the causes, effects, agent, stative, and theme of an event. The research on ATOMIC has drawn more and more attention in recent years. An increasing number of downstream tasks, including commonsense reasoning (Yu et al., 2022), storytelling (Brahman and Chaturvedi, 2020), question answering (Heo et al., 2022), dialog generation (Wu et al., 2022), etc., have improved their performances by acquiring and utilizing the commonsense knowledge from ATOMIC.

Currently, ATOMIC was constructed under one-hop annotations. It began with 24,000 pre-defined base events and nine relation types. For each base event and each relation, the annotators were asked to write a possible tail event based on one-hop reasoning. As shown in Figure 1, given the base event *"X asks Y to marry"*, the annotated tail events can be *"loving"* under the relation of *"xAttr"*, *"smiles"* under the relation of *"xEffect"*, and *"says yes"* under the relation of *"oEffect"*.

In such a one-hop annotation manner, each base event and its related annotated tail events shape a bipartite graph containing only $\mathcal{B}$-to-$\mathcal{A}$ links, where $\mathcal{B}$ denotes the **B**ase event and $\mathcal{A}$ denotes the **A**nnotated tail event. Thereby, the whole graph of ATOMIC can be viewed as a set of $\mathcal{B}$-to-$\mathcal{A}$ bipartite graphs, while the $\mathcal{B}$-to-$\mathcal{B}$, $\mathcal{A}$-to-$\mathcal{B}$ and $\mathcal{A}$-to-$\mathcal{A}$ links between different bipartite graphs were almost ignored. In Figure 1, the dashed lines illustrate such missing links in ATOMIC, e.g., an annotated tail event *"in front of Y"* and a base event *"X asks Y to marry"* in two different bipartite graphs miss a link of the *"xIntent"* relation.

This leads to two shortcomings of ATOMIC. Firstly, with only $\mathcal{B}$-to-$\mathcal{A}$ links, ATOMIC contains very few multi-hop paths, since an annotated tail event cannot become the *head event* of a triplet. Secondly, missing $\mathcal{B}$-to-$\mathcal{B}$, $\mathcal{A}$-to-$\mathcal{B}$ and $\mathcal{A}$-to-$\mathcal{A}$ links cause unsatisfactory knowledge coverage, despite its high-quality human-annotated commonsense knowledge. Both shortcomings limit the potential of ATOMIC in practical applications. Intuitively, an ideal CSKG requires high knowledge coverage to meet the needs of various tasks, and massive multi-hop paths to understand the evolu-

*Corresponding author

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 13292–13305
July 9-14, 2023 ©2023 Association for Computational Linguistics

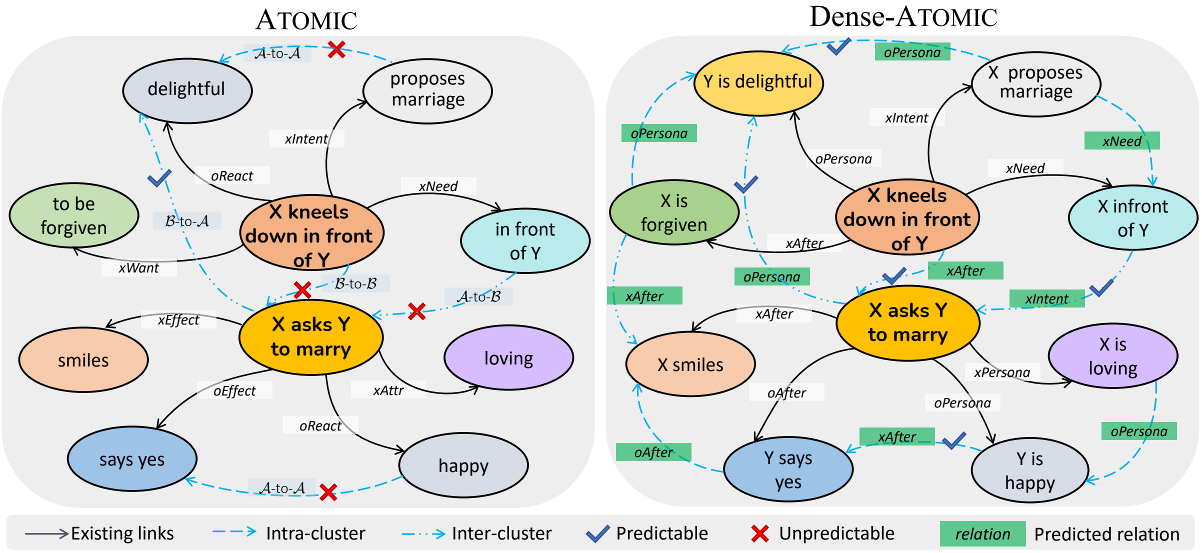

Figure 1: ATOMIC vs. Dense-ATOMIC. Firstly, Dense-ATOMIC completes many missing links in ATOMIC, including $\mathcal{B}$-to-$\mathcal{A}$, $\mathcal{B}$-to-$\mathcal{B}$, $\mathcal{A}$-to-$\mathcal{B}$, and $\mathcal{A}$-to-$\mathcal{A}$ links, *e.g.*, missing "oPersona" link between *"X proposes marriage"* and *"Y is delightful"* (type: $\mathcal{A}$-to-$\mathcal{A}$); Secondly, Dense-ATOMIC contains more multi-hop paths, e.g., a two-hop path *"X asks Y to marry"* → *"Y says yes"* → *"X smiles"* after predicting missing links on normalizd ATOMIC.

tion between different events.

In this work, we aim to construct a densely-connected ATOMIC. The key is to complete different types of missing links, leading to denser ATOMIC with high knowledge coverage and massive multi-hop paths. We achieve this goal through three main steps: Normalizing Tail Events, Training a Relation Prediction Model and Constructing Dense-ATOMIC.

Firstly, most of the annotated tail events in ATOMIC have different patterns to the base events, so we normalize annotated tail events in ATOMIC to a consistent pattern (*"Subject + Verb + Object"*), to facilitate subsequent CSKG completion. Specific relations are also grouped to mitigate ambiguity.

Secondly, we train a relation prediction model based on a set of existing triplets in ATOMIC to infer the missing links on the whole graph, *i.e.*, CSKG completion upon ATOMIC. To the best of our knowledge, most of the existing studies for CSKG completion utilized the translation based methods, which formalized the CSKG completion as a *tail event* ranking task given the *head event* and the relation. A graph convolutional network (GCN) was mostly employed to encode the graph embeddings of events, but its performance is unsatisfactory since the sparsity of ATOMIC limits the information propagation on the GCN (Malaviya et al., 2020). In contrast, in this work, we propose a method called Rel-CSKGC, which regards CSKG completion as a relation prediction problem given the *head event* and the *tail event*, and accordingly train a CSKG completion model based on ATOMIC.

Finally, based on the CSKG completion model, we construct Dense-ATOMIC by inferring the missing links on ATOMIC. Figure 1 illustrates the main differences between ATOMIC and Dense-ATOMIC.

We conduct extensive evaluations towards the Rel-CSKGC method and the constructed Dense-ATOMIC, respectively.

First, we compare Rel-CSKGC with several newly proposed relation prediction methods and translation based methods. Both automatic evaluation on an annotated subgraph and human evaluation on 500 sampled triplets show the advantage of Rel-CSKGC for completion on ATOMIC .

Next, we evaluate Dense-ATOMIC from the perspectives of knowledge coverage and multi-hop paths respectively. Extensive experiments are conducted in terms of statistics, human evaluation, and simple downstream tasks. The results demonstrate that Dense-ATOMIC surpasses ATOMIC in terms of triplet counts by an order of magnitude, and multi-hop paths by more than two orders of magnitude, respectively, while at the same time maintaining its quality.

2 Approach

Figure 2 illustrates the procedure of constructing Dense-ATOMIC, consisting of three main steps:

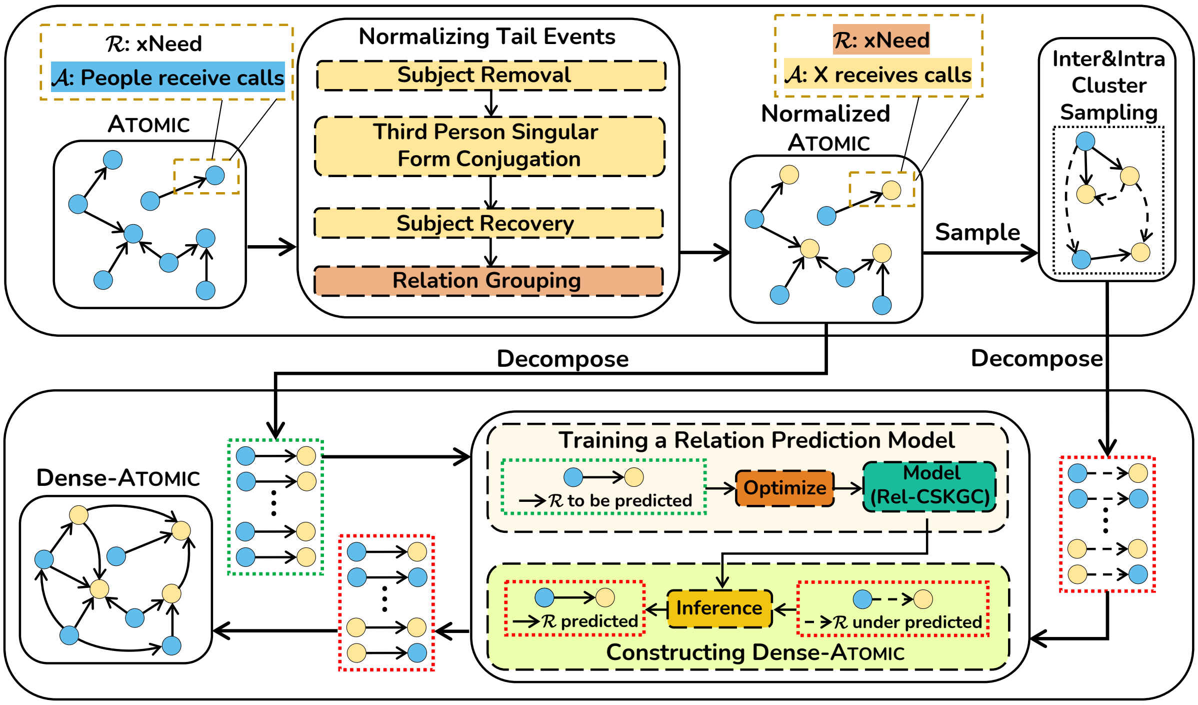
$\mathcal{R}$: xNeed
$\mathcal{A}$: People receive calls
ATOMIC
Normalizing Tail Events
Subject Removal
Third Person Singular Form Conjugation
Subject Recovery
Relation Grouping
$\mathcal{R}$: xNeed
$\mathcal{A}$: X receives calls
Normalized ATOMIC
Sample
Inter&Intra Cluster Sampling
Decompose
Decompose
Dense-ATOMIC
Training a Relation Prediction Model
$\mathcal{R}$ to be predicted
Optimize
Model (Rel-CSKGC)
$\mathcal{R}$ predicted
Inference
$\mathcal{R}$ under predicted
Constructing Dense-ATOMIC

Figure 2: Procedure of constructing Dense-ATOMIC.

Normalizing Tail Events, Training a Relation Prediction Model, and Constructing Dense-ATOMIC.

2.1 Normalizing Tail Events

ATOMIC contains only $\mathcal{B}$-to-$\mathcal{A}$ triplets. A CSKG completion model trained with $\mathcal{B}$-to-$\mathcal{A}$ triplets is inapplicable to predict $\mathcal{B}$-to-$\mathcal{B}$, $\mathcal{A}$-to-$\mathcal{A}$, and $\mathcal{A}$-to-$\mathcal{B}$ links, since base events (usually sentences) and annotated tail events (usually phrases or words) have different patterns. This results in a shortage of knowledge coverage and multi-hop paths during the completion.

To this end, we propose Normalizing Tail Events to convert annotated tail events to the same pattern as the base events, including subject removal, third person singular form conjugation, subject recovery, and relation grouping.

Subject Removal For a few annotated tail events being complete sentences, we perform dependency tree parsing and part-of-speech tagging with CoreNLP (Manning et al., 2014) and remove subjects based on the two kinds of structure patterns, which makes the nodes in the graph become a uniform pattern and benefits the subject recovery process. For example, given a tail event "He smiles", we first remove the subject "He" and convert it to a universal expression "Y smiles" in the subject recovery process.

Third Person Singular Form Conjugation In our preliminary experiments, a CSKG completion model tends to correlate phrases starting with *"to"* with relations such as *"xWant"*, *"xIntent"*, so we leverage WordNet (Miller, 1995) to acquire the verb root and add the suffix (-s, -es, etc.) according to English grammar.

Subject Recovery We add subjects to processed annotated tail events based on different relations.

Relation Grouping Both *"xWant"* and *"xEffect"* describe the possible subsequent events, distinguished by *"to"* representing subject will. After third person singular form conjugation, the two relations may lead to ambiguity. We perform relation grouping for all these relations to mitigate ambiguity. *"xEffect"* and *"xWant"* form *"xAfter"* describing *what will happen to X*. *"oEffect"* and *"oWant"* form *"oAfter"* describing *what will happen to Y*. *"xAttr"* and *"xReact"* form *"xPersona"* describing *how X feels or is described*. It should be noted that the relation grouping process leads to a non-serious problem, i.e., the grouped relation cannot distinguish between subjective and objective semantics. However, it mitigates ATOMIC's sparsity issue and improves the performance of the relation prediction model.

Due to the page limitation, the pseudo-code of normalizing tail events is present in Appendix A.

It is worth noting that our normalization method resembles a prior work (Fang et al., 2021b,a). Their purpose is to align ATOMIC with other CSKGs, while we focus on event alignment in ATOMIC by eliminating differences among different events.

2.2 Training a Relation Prediction Model

2.2.1 Limitation of Traditional Methods

Traditional methods for the completion of ATOMIC proposed to score all candidate *tail events* given the *head event* and the relation. The GCN for encoding graph embeddings of events induced two shortcomings: 1) it is difficult for a GCN to propagate information due to the sparse graph structure of ATOMIC (Malaviya et al., 2020); 2) it cannot sufficiently utilize semantic information of events.

2.2.2 Our Rel-CSKGC Method

To address these issues, we propose Rel-CSKGC, as illustrated in Figure 3. Specifically, ATOMIC is first decomposed into independent triplets, and then Rel-CSKGC predicts the relation given the *head event* and the *tail event* of a triplet. Rel-CSKGC utilizes no graph structure information thus avoiding the problem caused by the sparsity. Additionally, encoding both the *head event* and the *tail event* with the pretrained language model successfully takes advantage of semantic information.

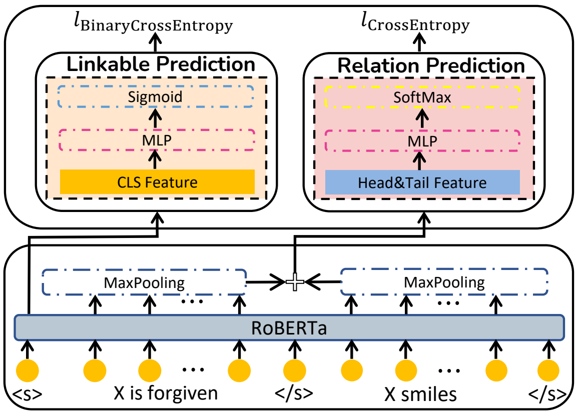

Figure 3: The detailed structure of Rel-CSKGC.

Problem Formulation Given a CSKG $G = (N, V)$, where N is the set of nodes and V is the set of edges, we consider a single training instance as a triplet $v_i = (h, r, t)$ with the *head event* h, *relation type* r and the *tail event* t. Here, $r \in V$ and $h, t \in N$. The objective of Rel-CSKGC is to predict the most reasonable r given h and t. [1]

Main Structure We utilize RoBERTa (Liu et al., 2019) to acquire contextual representations of free-form texts describing events. The input is the concatenation of h and t. We acquire the embedding matrix of h and t by:

$$[H; T] = \text{RoBERTa}([h; t]) \tag{1}$$

where $H \in \mathbb{R}^{|N| \times D}$ and $T \in \mathbb{R}^{|N| \times D}$. $|N|$ is the number of tokens of the event, and D is the dimensionality of representation. We apply max pooling on H and T to acquire sentence embeddings e_h and e_t. The objective function can be defined with trainable weights $W_t \in \mathbb{R}^{1 \times D}$ and $W_c \in \mathbb{R}^{K \times 2D}$:

$$o = \text{sigmoid}(W_t e_{\text{<s>}}) + \text{softmax}(W_c(e_h, e_t)) \tag{2}$$

where K is the number of relations and $e_{\text{<s>}}$ the embedding of <s>-token used as a indicator for whether h and t are related.

Negative Sampling Rel-CSKGC requires negative samples to predict *unlinkable* links. We consider the following two strategies to construct negative samples: 1) **Random** negative sampling. For a gold triplet, we randomly select an event from normalized ATOMIC as the new *tail event* to replace the original *tail event*; 2) **Persona** negative sampling. Triplets under relations of *"xPersona"* and *"oPersona"* follow the pattern of *"Subject + is + Adjective"* and account for a large part in ATOMIC. Models tend to always predict *"xPersona"* or *"oPersona"* when the given tail event follows the pattern of *"Subject + is + Adjective"*. To alleviate this problem, we specifically construct negative samples by replacing the *tail event* of triplets under relations of *"xPersona"* and *"oPersona"* with a randomly-chosen event containing "is".

2.3 Constructing Dense-ATOMIC

Based on Rel-CSKGC, we train a relation prediction model with existing triplets in ATOMIC and then use the model to complete missing links in ATOMIC. We adopt threshold-based link prediction to decide whether two events are related and propose an intra-and-inter cluster completion strategy to reduce the cost of completing entire ATOMIC.

Threshold-based Link Prediction Threshold-based link prediction (TLP) is a heuristic strategy to decide whether a relation is acceptable according to the probability predicted by Rel-CSKGC. Different thresholds are specifically tuned for different relations. The model predicts the relation

[1] To keep ATOMIC concise, we only predict the most reasonable relation in this work.

only if the final probability is above the corresponding threshold. TLP is used in all our models as the last step for the link acceptance decision.

Intra-and-inter Cluster Completion Strategy Since it's computationally expensive to iterate over all pairs of *head* and *tail event*s during the inference, we design an intra-and-inter cluster completion strategy to trade off between the completion scale and the time complexity. In Figure 1, we consider each base event and its annotated tail events as a *cluster*. **Intra-cluster completion** infers missing links inside a cluster. Intuitively, annotated tail events in one cluster, written based on the same base event, are highly related and may contain more missing links. **Inter-cluster completion** infers missing links between different clusters. Annotated tail events in different clusters are written independently based on different base events, thus links between different clusters are under-explored.

Due to the limited computing resource and time, we temporarily provide the results of 100 sampled clusters in this paper. Increasing the sampling size can further improve the scale of Dense-ATOMIC, but that will also linearly increases the computational cost. We will release versions with larger sampling sizes later.

3 Evaluation of Our Rel-CSKGC Method

In this section, we compare Rel-CSKGC with relation prediction and translation based methods by experimenting on a newly annotated subgraph and human evaluation.

3.1 Training and Test Set Construction

Training Set with Negative Sampling Following Sap et al. (2019)'s split of ATOMIC, we randomly sample negative triplets from the training split with negative sampling strategies introduced in Section 2.2. We combine sampled negative triplets and the training split to construct the training set for Rel-CSKGC. The statistic of the training set is illustrated in Table 1. [2]

ATOMIC	Rand. Neg. Samples	Per. Neg. Samples
463,264	1,890,350	756,140

Table 1: Statistics of the training set for Rel-CSKGC.

Test Set with Annotated Subgraph To test the performance of Rel-CSKGC, we construct a ground-truth subgraph by randomly sampling three clusters from the test split and annotating all pairs of *head event*s and *tail event*s with the most reasonable relation. The statistic of the annotated ground-truth subgraph is shown in Table 2.

Relation	Total	Intra	Inter
xAfter	243	186	57
xNeed	66	64	2
xIntent	72	51	21
xPersona	291	226	65
oAfter	262	174	88
oPersona	114	70	44
NoLink	4234	2303	1931

Table 2: Statistics of the annotated subgraph. Intra and Inter indicate the intra- and inter- cluster, respectively.

3.2 Compared Methods

We select 4 baselines comprising two different types of CSKG completion methods and use the specific evaluation protocol for each of them.

3.2.1 Relation Prediction Methods

Baselines We adapt **CE-random** (Li et al., 2016), a method augmenting CSKGs by scoring novel tuples, to predict the missing relation. We also compare **KG-BERT** (Yao et al., 2019), which probes the performance of relation prediction methods on knowledge graphs. Note that we replace BERT (Devlin et al., 2019) with RoBERTa (Liu et al., 2019) in KG-BERT for fair comparison.

Evaluation Protocal Ranking metrics (HITS and Mean Reciprocal Rank) designed for translation based methods are not applicable to relation prediction methods. By valuing precision more than recall on CSKG completion, we utilize precision for the evaluation of relation prediction methods.

3.2.2 Translation Based Methods

Baselines **SynLink** (Malaviya et al., 2020) proposed to densify the CSKG with synthetic links for better graph representation. **InductiveE** (Wang et al., 2021) introduced indutive learning on the CSKG by enhancing the unseen event representations with neighboring structure information.

Evaluation Protocal To handle the evaluation mismatch between Rel-CSKGC and translation

[2] The imbalance between random and persona negative sampling methods was established based on a preliminary experiment, which provided insights into optimal sampling sizes.

based methods, we designed a transformation strategy. Specifically, we randomly sample 500 triplets from Malaviya et al. (2020)'s test split. For SynLink and InductivE, a threshold is set for hit@1 score, and a *tail event* is accepted only when the score is above the threshold. We tune the threshold to ensure the number of triplets inferred by Rel-CSKGC, SynLink, and InductivE close on these 500 triplets. We then calculate the proportion of meaningful triplets for different methods manually.[3]

3.3 Main Results

Relation Prediction Methods In Table 3, we compare Rel-CSKGC with different relation prediction methods, and Rel-CSKGC achieves consistent improvement on the test set of the annotated subgraph. Paired t-Test result proves that the improvement of Rel-CSKGC is significant. From Table 3, we can observe that the precision of intra-cluster completion is significantly higher than that of inter-cluster completion for all methods. This demonstrates that tail events annotated based on the same base event are highly related to each other and easier for models to predict relations, while the prediction for inter-cluster events is more challenging.

Method	Total	Intra	Inter
CE-random	0.45	0.53	0.29
KG-BERT	0.60	0.67	0.43
Rel-CSKGC	**0.68**	**0.78**	**0.51**
- w/o random	0.36	0.45	0.22
- w/o persona	0.58	0.66	0.44
Rel-CSKGC$_{human}$	0.80	0.91	0.62

Table 3: Rel-CSKGC vs. Relation Prediction methods on Precision. Intra and Inter indicate the result of the intra- and inter- cluster, respectively.

Method	# Predicted	# Meaningful	Proportion
SynLink$_{Adapt}$	133	93	0.70
InductivE$_{Adapt}$	132	106	0.80
Rel-CSKGC	**174**	**152**	**0.87**

Table 4: Rel-CSKGC vs. Translation Based methods.

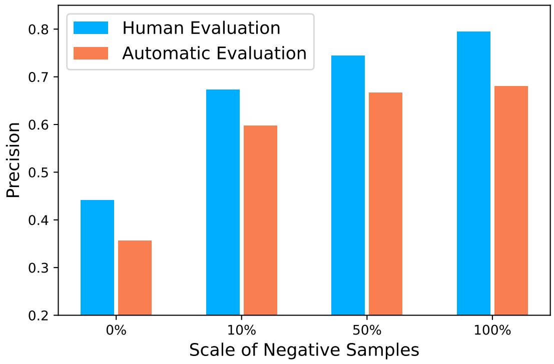

Figure 4: Precision of Rel-CSKGC with different scales of negative samples on the test set by automatic and human evaluation.

Translation Based Methods After carefully tuning the threshold based on the strategy in Section 3.2.2, Rel-CSKGC, SynLink, and InductivE predict 174, 133, and 132 triplets on 500 randomly sampled triplets. In Table 4, Rel-CSKGC outperforms SynLink and InductivE by a large margin on proportion and the number of meaningful triplets.

3.4 Human Evaluation

Motivation Upon observing predictions of Rel-CSKGC, we note that some triplets could be reasonable, while the annotated subgraph doesn't cover them. For example, given a *head event* "X accepts Y's apology" and a *tail event* "X is generous", the annotated ground-truth relation is "xPersona", while Rel-CSKGC could predict another reasonable relation "xIntent". Consequently, we perform the human evaluation to check whether a predicted triplet is actually meaningful.

Result We can find from the last row of Table 3 that Rel-CSKGC achieves an even higher precision of 0.80, suggesting that Rel-CSKGC can predict reasonable triplets neglected during the subgraph annotation. The high precision by human evaluation also guarantees the quality of predicted triplets.

3.5 Ablation Study

To validate the effectiveness of negative sampling, we report experimental results without negative sampling in Table 3. The performance of Rel-CSKGC drops dramatically without any negative sampling strategies, validating the effectiveness of negative sampling.

By experimenting Rel-CSKGC with different scales of random negative samples in Figure 4, we find that the precision of Rel-CSKGC increases using both automatic and human evaluation as more negative samples are used for training.

[3] In the given context, "meaningful triplets" refer to triplets that are considered reasonable, coherent, and non-contradictory by human evaluators.

4 Evaluation of the Constructed Dense-ATOMIC

4.1 Knowledge Coverage and Quality

In this subsection, we aim to answer the following question: *Does* Dense-ATOMIC *yield higher knowledge coverage while ensuring the quality?*

To this end, we statistically and manually compare Dense-ATOMIC with ATOMIC from the following three perspectives.

	# Events	# 1-hop	# 2-hop	# 3-hop
ATOMIC	299,068	696,321	19,231	509
Dense-ATOMIC	**283,435**	**1,967,373**	**10,658,242**	**67,888,373**

Table 5: ATOMIC vs. Dense-ATOMIC on the number of events and multi-hop paths.

Dense-ATOMIC yields higher knowledge coverage In Table 5, we present the comparison between ATOMIC and Dense-ATOMIC. Dense-ATOMIC contains 3x more one-hop paths than ATOMIC, contributing a significantly higher knowledge coverage. It's worth noting that different tail events in ATOMIC could become the same after normalizing tail events, so Dense-ATOMIC contains slightly fewer events than ATOMIC.

Triplets in Dense-ATOMIC have relatively high precision In Table 3, Rel-CSKGC achieves a precision of 0.80 by human evaluation. Moreover, from comparison results with translation based methods in Table 4, Rel-CSKGC outperforms two state-of-the-art methods by more than 7 percentage points. The high performance of Rel-CSKGC ensures the quality of predicted triplets to a certain extent.

Dense-ATOMIC benefits the performance of COMET To empirically demonstrate the knowledge coverage and quality of Dense-ATOMIC, we evaluate Dense-ATOMIC with COMET (Bosselut et al., 2019). The relation distribution of Dense-ATOMIC is long-tailed. We randomly sample 262,678 triplets from predicted triplets and recover the grouped relations to their original relations by following the relation distribution of the Sap et al. (2019)'s training split. Apart from the evaluation of perplexity, we design a strategy to evaluate the diversity score of generated *tail event*s. For each relation, we randomly sample 10 *head events* from the test set. For each test sample consisting of a *head event* and a relation, 10 candidates are generated using beam search. For each candidate, we manually give a score of 0, 1, or 2, representing "unreasonable", "plausible", and "reasonable", respectively. We then merge candidates of similar semantics into a group and calculate the group average score. The diversity score of 10 candidates is the sum of the group scores. Intuitively, the lower perplexity and the higher diversity score indicate the higher knowledge quality and the higher knowledge coverage of Dense-ATOMIC, and COMET$_{ours}$ outperforms COMET on both metrics in Table 6. In Table 7, we can find that tail events generated by COMET$_{ours}$ are more semantically different.

	PPL ↓	DS ↑
COMET	11.14	9.16
COMET$_{ours}$	**11.11**	**10.77**

Table 6: COMET vs. COMET$_{ours}$. PPL and DS indicate perplexity and diversity score, respectively.

COMET	COMET$_{ours}$
to study hard	**to study harder**
study hard	**to study more**
to study more	**to get a good grade**
to study	**to take a test**
to get a good grade	**to do well in school**
to take a test	**to do well in class**
to do well in school	**to apply for a job**
to get a good job	**to pass the class**
to apply for a job	**to get a prize**
to apply for a good job	**to go to school**

Table 7: Events generated by COMET and COMET$_{ours}$ given "***X needs a good grade***" and "***xWant***". Semantically similar events are in the same color.

4.2 Multi-hop Paths in Dense-ATOMIC

The aim of this subsection is to answer the question: *Can multi-hop paths in Dense-*ATOMIC *better present the commonsense knowledge?*

Accordingly, we evaluate multi-hop paths based on the human evaluation and performing a newly designed Commonsense Reasoning experiment, respectively:

Sampling Method	2-hop	3-hop	4-hop
Random	0.69	0.62	0.50
Heuristic Rule	**0.84**	**0.77**	**0.74**

Table 8: Random vs. Heuristic Rule on human evaluation of sampled multi-hop paths.

2-hop paths
X misses Y's opportunity $\xrightarrow{xAfter}$ X goes home sadly $\xrightarrow{xPersona}$ X is melancholy
X takes advantage of the opportunities $\xrightarrow{xAfter}$ X contines to succeed $\xrightarrow{oPersona}$ Y is satisfied
X goes back home $\xrightarrow{xAfter}$ X becomes sleepy $\xrightarrow{xAfter}$ X goes back to his own bed
X reaches X's goal $\xrightarrow{xAfter}$ X gets an award $\xrightarrow{oAfter}$ Y celebrates their win
3-hop paths
X returns to X's work $\xrightarrow{xAfter}$ X goes home for the day $\xrightarrow{xAfter}$ X sleeps at night $\xrightarrow{oAfter}$ Y is glad to see X slept normally
X plays a role in the development $\xrightarrow{xAfter}$ X receives an award $\xrightarrow{xAfter}$ X gets compliments $\xrightarrow{xAfter}$ X smiles
X talkes about X's feeling $\xrightarrow{xAfter}$ X starts crying $\xrightarrow{xAfter}$ X wipes the tears $\xrightarrow{xPersona}$ X is thankful
X improves X's chances $\xrightarrow{xAfter}$ X wins the game $\xrightarrow{xAfter}$ X jumps up and down with joy $\xrightarrow{oPersona}$ Y is pleased

Table 9: Examples of multi-hop paths randomly sampled from Dense-ATOMIC.

Human evaluation confirms the correctness of multi-hop paths in Dense-ATOMIC In Table 5, we have already shown that Dense-ATOMIC contains orders of magnitude more two-hop and three-hop paths than ATOMIC. Now, to further validate the correctness of multi-hop paths, we perform the human evaluation on sampled paths to calculate the proportion of reasonable paths. Note that it's a common phenomenon (both KGs and CSKGs) that $A \rightarrow B$ and $B \rightarrow C$ are reasonable, while $A \rightarrow B \rightarrow C$ is irrational. For example, {*Beethoven*, *owner*, *piano*} and {*piano*, *color*, *black*} are two reasonable triplets, but "*Beethoven*" and "*black*" are not related. Consequently, we additionally design a simple heuristic sampling rule: a multi-hop path $A \rightarrow \ldots \rightarrow C$ is chosen only when A and C are also linked in Dense-ATOMIC. By comparing with random sampling in Table 8, we can find that heuristic rule sampling consistently outperforms random sampling: the longer the multi-hop paths, the more significant the improvement. Multi-hop paths randomly sampled from Dense-ATOMIC with two different methods are illustrated in Table 9.

Dense-ATOMIC has the potential of providing contextual information for Commonsense Reasoning In order to further validate the effectiveness of multi-hop paths in Dense-ATOMIC, we utilize BART (Lewis et al., 2020) to perform generative Commonsense Reasoning with or without multi-hop paths. Specifically, with the heuristic rule above, we randomly sample 5000 four-hop paths from Dense-ATOMIC as the training samples. For test samples, we manually select 500 reasonable paths from Dense-ATOMIC. BART is trained to generate the subsequent event in two different settings: 1) given only the first node of the path; 2) given the first four nodes of the path. From Table 10, we can find that BART trained with multi-hop paths achieves better performance in that multi-hop paths could provide more contextual information useful for Commonsense Reasoning.

	Bleu-1	Bleu-2	ROUGE-L
One-hop	48.57	14.24	35.58
Multi-hop	**48.63**	**14.93**	**36.90**

Table 10: Scores of tail events generated with one-hop and multi-hop paths.

5 Related Work

ConceptNet (Speer et al., 2017) is a large-scale CSKG merging various knowledge bases. ASER (Zhang et al., 2020b) contains the selectional preference knowledge extracted from more than 11 billion-token unstructured textual data. TransOMCS (Zhang et al., 2020a) utilizes linguistic graphs to convert ASER into the same representation as ConceptNet. DISCOS (Fang et al., 2021b) aggregates the neighboring information to distill the commonsense knowledge in ASER.

Recent years have seen crowdsourced CSKGs aiming to provide high-quality commonsense knowledge triplets. Sap et al. (2019) released ATOMIC consisting of if-then knowledge triplets mainly about daily events. Hwang et al. (2021) augmented ATOMIC with event-centered and physical-entity triplets. GLUCOSE (Mostafazadeh et al., 2020) grounds the implicit commonsense knowledge about everyday situations in a narrative context for richer inferential content.

Dense-ATOMIC unleashes the power of ATOMIC for high knowledge coverage and multi-hop paths.

Prior CSKG completion methods performed binary classification by scoring BiLSTM-encoded tuples (Li et al., 2016; Saito et al., 2018; Jastrzębski et al., 2018). Following translation based methods for the knowledge graph completion (Dettmers et al., 2018; Shang et al., 2019; Meilicke et al., 2019; Qu et al., 2021; Zhang et al., 2021; Lovelace et al., 2021), Malaviya et al. (2020) additionally densified the CSKG based on BERT similarity and achieve promising results. Wang et al. (2021) and Ju et al. (2022) designed heuristic rules to add more edges for nodes with fewer neighbors. Moghimifar et al. (2021) presented a neural-symbolic reasoner to learn logic rules during the training, making the CSKG completion process interpretable.

Rel-CSKGC differs from them in that we utilize pretrained language models to predict the relation given the *head event* and the *tail event*. Similar relation prediction methods targeting at the knowledge graph completion have been proposed (Socher et al., 2013; Yao et al., 2019; Cao et al., 2020). To our best knowledge, we are the first to explore the relation prediction method on CSKG completion.

6 Conclusion

In this paper, we construct Dense-ATOMIC for high knowledge coverage and massive multi-hop paths and accordingly propose a CSKG completion method called Rel-CSKGC to train a relation prediction model and infer the missing links in ATOMIC. Both automatic and human evaluation show the advantage of Rel-CSKGC over strong baselines. The statistics prove that Dense-ATOMIC has significantly more triplets and multi-hop paths, providing potential for high-quality downstream applications and multi-hop reasoning based on commonsense knowledge.

Limitations

Our approach for constructing Dense-ATOMIC still has two limitations: 1) to keep Dense-ATOMIC simple, we only consider the most reasonable relation in this paper, while the relation between two events can be complex and diversified. We will release versions of Dense-ATOMIC with diversified relations later; 2) due to page limitation, we only evaluate Dense-ATOMIC on simple commonsense reasoning tasks, and we will further validate the multi-hop reasoning capacity of Dense-ATOMIC on more complex downstream tasks in the future.

Ethics Statement

We would like to thank the Allen Institute for AI for their valuable work on ATOMIC. The ATOMIC is licensed under a license of CC BY, which allows remixing, transforming, and building upon the material for any purpose. We will also make our Dense-ATOMIC publicly available later. Mehrabi et al. (2021) have found representational harms in common sense resources. We acknowledge that the generated commonsense from our models might contain biases. All of the datasets and models are in English, which benefits English speakers more. We have employed 3 postgraduates experienced in natural language processing for annotation and human evaluation. We pay postgraduates around $8 per hour, well above the local average wage, and engage in constructive discussions if they have concerns about the process.

Acknowledgments

This work was supported by the Natural Science Foundation of China (No. 62076133), and the Natural Science Foundation of Jiangsu Province for Distinguished Young Scholars (No. BK20200018).

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4762–4779. Association for Computational Linguistics.

Faeze Brahman and Snigdha Chaturvedi. 2020. Modeling protagonist emotions for emotion-aware storytelling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5277–5294. Association for Computational Linguistics.

Ermei Cao, Difeng Wang, Jiacheng Huang, and Wei Hu. 2020. *Open Knowledge Enrichment for Long-Tail Entities*, page 384–394. Association for Computing Machinery, New York, NY, USA.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1811–1818. AAAI Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Tianqing Fang, Weiqi Wang, Sehyun Choi, Shibo Hao, Hongming Zhang, Yangqiu Song, and Bin He. 2021a. Benchmarking commonsense knowledge base population with an effective evaluation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8949–8964. Association for Computational Linguistics.

Tianqing Fang, Hongming Zhang, Weiqi Wang, Yangqiu Song, and Bin He. 2021b. DISCOS: bridging the gap between discourse knowledge and commonsense knowledge. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 2648–2659. ACM / IW3C2.

Yu-Jung Heo, Eun-Sol Kim, Woo Suk Choi, and Byoung-Tak Zhang. 2022. Hypergraph transformer: Weakly-supervised multi-hop reasoning for knowledge-based visual question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 373–390. 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 *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 6384–6392. AAAI Press.

Stanislaw Jastrzębski, Dzmitry Bahdanau, Seyedarian Hosseini, Michael Noukhovitch, Yoshua Bengio, and Jackie Cheung. 2018. Commonsense mining as knowledge base completion? a study on the impact of novelty. In *Proceedings of the Workshop on Generalization in the Age of Deep Learning*, pages 8–16, New Orleans, Louisiana. Association for Computational Linguistics.

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, Atlanta, GA, USA, October 17-21, 2022*, pages 4104–4108. ACM.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin 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, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

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.

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. *CoRR*, abs/1907.11692.

Justin Lovelace, Denis Newman-Griffis, Shikhar Vashishth, Jill Fain Lehman, and Carolyn P. Rosé. 2021. Robust knowledge graph completion with stacked convolutions and a student re-ranking network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1016–1029. Association for Computational Linguistics.

Chaitanya Malaviya, Chandra Bhagavatula, Antoine Bosselut, and Yejin Choi. 2020. Commonsense knowledge base completion with structural and semantic context. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 2925–2933. AAAI Press.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, System Demonstrations*, pages 55–60. The Association for Computer Linguistics.

Ninareh Mehrabi, Pei Zhou, Fred Morstatter, Jay Pujara, Xiang Ren, and Aram Galstyan. 2021. Lawyers are dishonest? quantifying representational harms in commonsense knowledge resources. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5016–5033. Association for Computational Linguistics.

Christian Meilicke, Melisachew Wudage Chekol, Daniel Ruffinelli, and Heiner Stuckenschmidt. 2019. Anytime bottom-up rule learning for knowledge graph completion. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 3137–3143. ijcai.org.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Farhad Moghimifar, Lizhen Qu, Terry Yue Zhuo, Gholamreza Haffari, and Mahsa Baktashmotlagh. 2021. Neural-symbolic commonsense reasoner with relation predictors. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 797–802. Association for Computational Linguistics.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David W. Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: generalized and contextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4569–4586. Association for Computational Linguistics.

Meng Qu, Junkun Chen, Louis-Pascal A. C. Xhonneux, Yoshua Bengio, and Jian Tang. 2021. Rnnlogic: Learning logic rules for reasoning on knowledge graphs. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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, Brussels, Belgium. Association for Computational Linguistics.

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*, 01, pages 3027–3035.

Chao Shang, Yun Tang, Jing Huang, Jinbo Bi, Xiaodong He, and Bowen Zhou. 2019. End-to-end structure-aware convolutional networks for knowledge base completion. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3060–3067. AAAI Press.

Richard Socher, Danqi Chen, Christopher D. Manning, and Andrew Y. Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 926–934.

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, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Bin Wang, Guangtao Wang, Jing Huang, Jiaxuan You, Jure Leskovec, and C.-C. Jay Kuo. 2021. Inductive learning on commonsense knowledge graph completion. In *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*, pages 1–8. IEEE.

Sixing Wu, Ying Li, Dawei Zhang, and Zhonghai Wu. 2022. KSAM: infusing multi-source knowledge into dialogue generation via knowledge source aware multi-head decoding. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 353–363. Association for Computational Linguistics.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. KG-BERT: BERT for knowledge graph completion. *CoRR*, abs/1909.03193.

Wenhao Yu, Chenguang Zhu, Lianhui Qin, Zhihan Zhang, Tong Zhao, and Meng Jiang. 2022. Diversifying content generation for commonsense reasoning with mixture of knowledge graph experts. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1896–1906. Association for Computational Linguistics.

Hongming Zhang, Daniel Khashabi, Yangqiu Song, and Dan Roth. 2020a. Transomcs: From linguistic graphs to commonsense knowledge. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4004–4010. ijcai.org.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020b. ASER: A large-scale eventuality knowledge graph. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 201–211. ACM / IW3C2.

Yao Zhang, Hongru Liang, Adam Jatowt, Wenqiang Lei, Xin Wei, Ning Jiang, and Zhenglu Yang. 2021. GMH: A general multi-hop reasoning model for KG completion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3437–3446. Association for Computational Linguistics.

A Algorithm for Normalizing Tail Events

Algorithm 1 presents the pseudo-code of Normalizing Tail Events in Section 2.1.

Algorithm 1 Normalizing Tail Events

```
Input: A set of annotations $A$ and relations $R$
Output: A set of sentences in present tense $FA$
Remove annotations with underscores or none, and get a series of filtered annotations $FA$
for each $fa \in FA$, $r \in R$ do
    Obtain the dependency tree $dep$ and POS tagging result $pos$ of $fa$
    Find $sub$ node with POS $prp$ and edge $subj$ connected directly to it
    if Position of $sub$ is at the start of $fa$ then
        Remove $sub$ in $fa$
    end if
    Find node $verb$ with POS $vb$ in $fa$
    if $r \in [xIntent, xWant, xNeed, oWant]$ AND the first word of $fa$ is *to* then
        Remove the first *to* of $fa$
    end if
    Transform node $verb$ in $fa$ to its root form
    Append $suf \in [-s, -es, -ies, ...]$ to $verb$ based on English grammar
    if $r \in [xAttr, xReact]$ then
        Insert *PersonX is* to the start of $fa$
    else if $r$ is $oReact$ then
        Insert *PersonY is* to the start of $fa$
    else if $r \in [oWant, oEffect]$ then
        Insert *PersonY* to the start of $fa$
    else
        Insert *PersonX* to the start of $fa$
    end if
end for
Return $FA$
```

B Implementation Details

Rel-CSKGC We use RoBERTa-large containing 335M parameters as the base model. We use a maximum sequence length of 100 and batch size of 128. The Adam optimizer is used for optimization with a learning rate of 2e-5 for RoBERTa-large and a learning rate of 1e-4 for MLP layers. The warmup proportion is set to 0.1. We train Rel-CSKGC with 1 NVIDIA RTX 3090 Graphical Card for 5 epochs, and it takes 20 hours to finish the training.

COMET$_{ours}$ To train COMET$_{ours}$, we use the implentations provided here. [4] We use the learning rate of 1.625e-5 and the default values for other parameters.

Generative Commonsense Reasoning BART-base is employed as the base model, which contains 140M parameters. We use a batch size of 128 and use the default values for other parameters.

[4] https://github.com/atcbosselut/comet-commonsense

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
In Limitations section.

□ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
In Abstract section and section 1, respectively.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

In section 2, Appendix B.

☑ B1. Did you cite the creators of artifacts you used?
In section 2.1, Appendix B.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
In Ethics Statement section.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
In Ethics Statement section.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We use publically available datasets, and the authors of the dataset have made the corresponding declaration.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
The documentation of the artifacts will be released after the reviewing process.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
In Section 3 and 4.

C ☑ Did you run computational experiments?

In Section 3 and 4.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
In Appendix B.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In Appendix B.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
In Section 3.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
In Appendix B.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

In Section 3 and 4.

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
We perform simple human annotation and evaluation, there is no need of providing the full text.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
In Ethics Statement.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
In Ethics Statement.