BUCA: A Binary Classification Approach to Unsupervised Commonsense Question Answering

Jie He[1] and **Simon Chi Lok U**[1] and **Víctor Gutiérrez-Basulto**[2] and **Jeff Z. Pan**[1]
[1] ILCC, School of Informatics, University of Edinburgh, UK
[2] School of Computer Science and Informatics, Cardiff University, UK
j.he@ed.ac.uk, c.l.u@sms.ed.ac.uk
gutierrezbasultov@cardiff.ac.uk, j.z.pan@ed.ac.uk

Abstract

Unsupervised commonsense reasoning (UCR) is becoming increasingly popular as the construction of commonsense reasoning datasets is expensive, and they are inevitably limited in their scope. A popular approach to UCR is to fine-tune language models with external knowledge (e.g., knowledge graphs), but this usually requires a large number of training examples. In this paper, we propose to transform the downstream multiple choice question answering task into a simpler binary classification task by ranking all candidate answers according to their reasonableness. To this end, for training the model, we convert the knowledge graph triples into reasonable and unreasonable texts. Extensive experimental results show the effectiveness of our approach on various multiple choice question answering benchmarks. Furthermore, compared with existing UCR approaches using KGs, ours is less data hungry. Our code is available at https://github.com/probe2/BUCA

1 Introduction

Commonsense reasoning has recently received significant attention in NLP research (Bhargava and Ng, 2022), with a vast amount of datasets now available (Levesque, 2011; Gordon et al., 2012; Sap et al., 2019; Rashkin et al., 2018; Bisk et al., 2020; Talmor et al., 2019). Most existing methods for commonsense reasoning either fine-tune large language models (LMs) on these datasets (Lourie et al., 2021) or use knowledge graphs (KGs) (Pan et al., 2017) to train LMs (Liu et al., 2019a; Yasunaga et al., 2022). However, it is not always possible to have relevant training data available, it is thus crucial to develop unsupervised approaches to commonsense reasoning that do not rely on labeled data.

In this paper, we focus on the unsupervised multiple choice question answering (QA) task: given a question and a set of answer options, the model is expected to predict the most likely option. We propose **BUCA**, a binary classification framework for unsupervised commonsense QA. Our method roughly works as follows: we first convert knowledge graph triples into textual form using manually written templates, and generate positive and negative question-answer pairs. We then fine-tune a pre-trained language model, and leverage contrastive learning to increase the ability to distinguish reasonable from unreasonable ones. Finally, we input each question and all options of the downstream commonsense QA task into BUCA to obtain the reasonableness scores and select the answer with the highest reasonableness score as the predicted answer. Experimental results on various commonsense reasoning benchmarks show the effectiveness of our proposed BUCA framework. Our main contributions are:

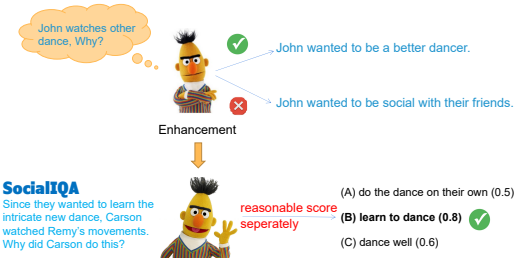

Figure 1: After BUCA is trained on the above question from the training set, it is then able to rate the reasonableness of each sentence of the downstream task.

- We propose a binary classification approach to using KGs for unsupervised commonsense question answering.
- We conduct extensive experiments, showing the effectiveness of our approach by using much less data.

2 Related work

Language models are widely used in unsupervised commonsense inference tasks, e.g. as an additional knowledge source or as a scoring model. Rajani

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 376–387

et al. (2019) propose an explanation generation model for the CommonsenseQA dataset. Self-talk (Shwartz et al., 2020) uses prompts to stimulate GPT and generate new knowledge. SEQA (Niu et al., 2021) generates several candidate answers using GPT2 and then ranks each them.

Another research direction in unsupervised commonsense reasoning is the use of e.g. commonsense KGs (Speer et al., 2016; Romero et al., 2019; Malaviya et al., 2020) to train the model (Chen et al., 2021; Geng et al., 2023). In Banerjee and Baral (2020), given the inputs of context, question and answer, the model learns to generate one of the inputs given the other two. Ma et al. (2021) update the model with a margin ranking loss computed on positive and negative examples from KGs. MICO (Su et al., 2022) uses the distance between the positive and negative question-answer pairs obtained from the KG to calculate the loss. However, all of the above approaches demand a large amount of training data, sometimes reaching million of training samples, while BUCA only needs tens of thousands, cf. Table 2. The most similar to our work is NLI-KB (Huang et al., 2021), which trains a model on NLI data, then applies the corresponding knowledge to each question-answer pair on the downstream task. Our paper, instead, shows that is not the NLI data but the retrieved knowledge that helps.

3 Methodology

We focus on the following multiple choice question answering (QA) task: given a question q and a set of options A, the model should select the most likely single answer $A_i \in A$. We consider an unsupervised setting in which the model does not have access to the training or validation data. Our BUCA approach first trains the model with a knowledge graph and then uses the trained model to test on multiple QA downstream tasks. Formally, a *knowledge graph (KG)* (Pan et al., 2017) $\mathcal{G}$ is a tuple (V, R, T), where V is a set of entities, E is a set of relation types and T is a set of triples of the form (h, r, t) with $h, t \in V$ the *head* and *tail* entities and $r \in R$ the *relation* of the triple connecting h and t.

Our approach has three main components: knowledge graph transfer to training data, training loss design, and downstream task testing:

Converting Triples into Binary Classification Training Data. Inspired by previous work (Su et al., 2022), each KG triple is converted into question-answer pairs by using pre-defined templates, so that the obtained pairs are then used as the input of the classification task. We use the templates provided in (Hwang et al., 2020). For example, the ATOMIC triple *(PersonX thanks PersonY afterwards, isAfter, PersonX asked PersonY for help on her homework)* can be converted to "*After PersonX asked PersonY for help on her homework, PersonX thanks PersonY afterwards*". In the appendix we show the distribution of the converted sequence pairs. Along with the correct QA pairs created from the KG triples, our framework is also trained on negative QA pairs, so it can better discriminate between reasonable and unreasonable QA pairs. More precisely, in the training dataset, each correct QA pair generated from a triple $tp = (h, r, t)$ has a corresponding negative pair obtained from a variation of tp in which t is substituted by t', which is randomly drawn from the existing tails in the KG.

Training Loss. For our binary classification model, we add a classification head with two nodes to the pre-trained language model. After normalizing the values on these two nodes, we can obtain reasonable and unreasonable scores for the QA pairs. From the triple conversion step, we obtained n training examples, each consisting of a question q, correct answer a_c, and incorrect answer a_w. For each question-answer pair, we can then obtain the reasonable and unreasonable scores r_i^+ and r_i^- after applying a softmax layer. In each loss calculation, we jointly consider the correct and incorrect answers. For binary classification, we use two kinds of losses: *Traditional Binary Loss (TBL).*

$$\mathcal{L} = -\sum_{i=1}^{n}(log(p_{a_c}^+) + log(p_{a_w}^-))$$

where $p_{a_c}^+$ and $p_{a_w}^-$ are the probabilities of correct and incorrect answers, respectively corresponding to reasonable and unreasonable scores.
Margin Ranking Loss.

$$\mathcal{L} = \sum_{i=1}^{n} max(0, \eta - log(p_{a_c}^+) + log(p_{a_w}^+)) + max(0, \eta - log(p_{a_w}^-) + log(p_{a_c}^-))$$

where η is a margin threshold hyper-parameter.

In order to pull the representational distance between reasonable question-answer pairs as close as possible and to push the representational distance

Methods	Backbone	Knowledge Source	COPA		OpenbookQA		SIQA	CSQA	SCT
			dev	test	dev	test	dev	dev	dev
Random	-	-	50.0	50.0	25.0	25.0	33.3	25.0	50.0
RoBERTa-L	RoBERTa-L	-	54.8	58.4	31.2	31.6	39.7	31.2	65.0
GPT2-L	GPT2-L	-	62.4	63.6	31.2	29.4	42.8	40.4	66.7
Self-talk	GPT2	GPT2	66.0	-	28.4	30.8	46.2	32.4	-
Dou	ALBERT	ALBERT	-	-	41.6	39.8	44.1	50.9	-
Wang	GPT2	GPT2	69.8	-	-	-	47.3	-	71.6
SMLM	RoBERTa-L	e.g., ATOMIC	-	-	34.6	33.8	48.5	38.8	-
MICO	RoBERTa-L	Concept	73.2	75.2	-	-	44.6	51.0	-
MICO	RoBERTa-L	ATOMIC	79.4	77.4	-	-	56.0	44.2	-
NLI-KB	RoBERTa-L	Concept	65.0	62.2	35.0	35.6	46.9	49.0	71.2
NLI-KB	RoBERTa-L	ATOMIC	65.2	61.6	39.0	37.2	46.7	52.1	72.1
Ma	RoBERTa-L	CSKG	-	-	-	-	63.2	67.4	-
BUCA	RoBERTa-L/TBL	Concept	84.4	**90.6**	43.0	47.2	53.5	63.5	87.3
BUCA	RoBERTa-L/MRL	Concept	**86.2**	89.6	45.2	**47.6**	52.6	**65.4**	88.0
BUCA	RoBERTa-L/TBL	ATOMIC	85.0	86.0	**45.8**	44.2	60.2	58.7	**88.4**
BUCA	RoBERTa-L/MRL	ATOMIC	84.6	87.8	43.2	46.0	**61.4**	60.3	85.5

Table 1: Accuracy (%) on five public benchmarks. Our best scores are highlighted in **bold**, and the results for the best performing baseline are underlined. Recall that TBL and MRL refer to the loss functions used in BUCA.

between reasonable and unreasonable ones as far as possible, we use supervised contrastive learning (Gunel et al., 2021) along with the binary classification. This is done by considering as positive examples of a given example within a category, all those examples within the same category.

Contrastive Loss of the i-th QA pair

$$\mathcal{L}_{scl} = \sum_{j=1}^{N} 1_{y_i=y_j} log \frac{e^{sim(h_j,h_i)\tau}}{\sum_{k=1}^{N} 1_{i\neq k} e^{sim(h_k,h_i)/\tau}}$$

where τ is the temperature parameter and h denotes the feature vector.

Inference. In the prediction phase for each candidate answer, we calculate its reasonableness score. We choose the answer with the highest reasonableness score as the predicted answer.

4 Experiments

In this section, we first describe our experiments on five commonsense question answering datasets, followed by ablation studies and data analysis.

4.1 Datasets and Baselines

We use two well-known commonsense KGs for training our framework: ConceptNet (Speer et al., 2017) and ATOMIC (Sap et al., 2018). For evaluation, we use five commonsense QA datasets: COPA (Gordon et al., 2012), OpenBookQA (Mihaylov et al., 2018), SIQA (Sap et al., 2019), CSQA (Talmor et al., 2019), and SCT (Mostafazadeh et al., 2017), covering a wide range of topics within commonsense reasoning. We compare our approach with various baselines: RoBERTa-Large (Liu et al., 2019b), GPT2 (Radford et al., 2019), Self-talk (Shwartz et al., 2020), Dou (Dou and Peng, 2022), Wang (Wang and Zhao, 2022) and other unsupervised systems using KGs: SMLM (Banerjee and Baral, 2020), MICO (Su et al., 2022), NLI-KB (Huang et al., 2021) and Ma (Ma et al., 2021). Most reported results are collected from the literature. For NLI-KB, we used their publicly available code to get the results.

Details of the KGs and datasets, as well as implementation details, can be found in the appendix.

Methods	Dataset	Train Pair	Valid Pair
Ma	ConceptNet	363,646	19,139
Ma	ATOMIC	534,834	60,289
Ma	WikiData	42,342	2,229
Ma	WordNet	256,922	13,523
MICO	WordNet	256,922	13,523
MICO	ATOMIC	1,221,072	48,710
BUCA	ConceptNet	65,536	7,836
BUCA	ATOMIC	61,053	2,435

Table 2: Statistics for the training and validation data used by Ma, MICO and BUCA.

4.2 Main results

Table 1 shows the results for the five benchmarks. Overall, BUCA achieves the best performance on all datasets. More precisely, our results respectively outperform baselines on the validation and test sets as follows: MICO by 6.8% and 13.2% on COPA; Dou by 4.2% and 7.8% on OpenbookQA. We also outperform MICO by 5.4% on SIQA; NLI-KB by 13.3% on CSQA, and NLI-KB by 16.3% on SCT. Ma does not provide results for COPA,

Backbone	CKG	COPA		OpenbookQA		SIQA	CSQA	SCT
		dev	test	dev	test	dev	dev	dev
BERT-base	Concept	63.0	67.6	29.6	32.8	40.5	49.6	64.9
BERT-base	ATOMIC	64.8	73.2	31.2	34.0	45.0	45.3	68.7
RoBERTa-base	Concept	70.0	72.8	30.0	32.8	46.6	49.0	65.6
RoBERTa-base	ATOMIC	70.4	77.4	33.4	34.2	50.6	46.9	70.6
RoBERTa-large	Concept	86.2	89.6	45.2	47.6	52.6	65.4	88.0
RoBERTa-large	ATOMIC	84.6	87.8	43.2	46.0	61.4	60.3	85.5

Table 3: Backbone model study

Backbone	CKG	COPA		OpenbookQA		SIQA	CSQA	SCT
		dev	test	dev	test	dev	dev	dev
RoBERTa-large	Concept	86.2	89.6	45.2	47.6	52.6	65.4	88.0
w/o contrastive	Concept	83.3	89.0	42.6	46.8	51.9	64.5	87.0
RoBERTa-large	ATOMIC	84.6	87.8	43.2	46.0	61.4	60.3	85.5
w/o contrastive	ATOMIC	84.2	86.6	42.0	44.0	60.6	59.8	84.1

Table 4: The influence of contrastive learning

OpenBookQA and SCT, but it achieves state-of-the-art results on CSQA 67.4 and on SIQA 63.2, while BUCA's best results respectively are 65.4 and 61.4. However, Ma uses multiple KGs to train a single model, ConceptNet, WordNet, and Wikidata for CSQA and ATOMIC, ConceptNet, WordNet, and Wikidata for SIQA, with a total training data of 662,909 and 1,197,742, while BUCA only uses 65,536 and 61,530, cf. Table 2. Considering the difference on used training data and the closeness of results, BUCA's approach clearly demonstrates its effectiveness. We can also observe the same trend as in MICO: ConceptNet is more helpful for CSQA and ATOMIC is more helpful for SIQA. This is explained by the fact that SIQA is built based on ATOMIC and CSQA is built based on ConceptNet. On other datasets our framework shows similiar behavior with both KGs. As for the loss functions, the margin ranking loss is on average 0.8% higher than the binary loss on ConceptNet, and 0.1% higher on ATOMIC. These results are explained by the fact that the ranking loss separates more the scores between reasonable and unreasonable answers. In light of this, we will only consider margin ranking loss in the below analysis.

4.3 Ablation Studies

In this section, we analyze the effects of the backbone models, the effect of contrastive learning, and explore the vocabulary overlap between the knowledge training set and the downstream task as well as the accuracy of our BUCA method.

Backbone Pre-trained LMs Our experiments using different backbone models show that in general the stronger the PLM the better the performance on the downstream task. Regarding the KGs, in the BERT-base and RoBERTa-base variants, the ATOMIC-trained models perform better than the ConceptNet-trained models, while in the RoBERTa-large one they perform similarly. This might be explained by the fact that as the model capacity increases it has more inherently available event-like commonsense knowledge, necessary in the ATOMIC-based datasets. Detailed results are shown in Table 3.

Effects of Contrastive Learning Our experiments show that the RoBERTa-large variant with contrastive learning outperforms the version without it on all datasets, regardless of the used KG. Detailed results are shown in Table 4.

Accuracy of the Binary Classifier Inspired by Ghosal et al. (2022), we evaluate how often input sequences corresponding to correct and incorrect answers are accurately predicted. To this end, we use the RoBERTa-large variant trained on ATOMIC. Table 5 shows that our model tends to predict all answers as reasonable since in our training set the negative examples are randomly selected, many QA pairs are semantically irrelevant or even ungrammatical. For the manually crafted candidate answers, many of them are semantically relevant and grammatical, so our model predicts them as reasonable. We also see that the accuracy metrics for SCT and COPA are the highest. Our findings are consistent with Ghosal et al. (2022).

4.4 Data Analysis

To better understand why transfer learning from CKGs is more suitable than from other datasets

Dataset	Prediction All				
	Neg	Pos	Incor as Neg	Cor as Pos	Accurate
COPA (dev)	0.2	88.0	11.2	99.0	11.0
COPA (test)	0.4	88.4	11.2	99.2	10.8
OpenbookQA (dev)	1.4	67.8	4.8	93.2	3.4
OpenbookQA (test)	1.8	73.8	2.8	93.0	1.0
SIQA (dev)	6.3	50.2	15.7	86.7	9.4
CSQA (dev)	1.2	35.1	6.5	94.2	5.2
SCT (dev)	0.3	87.8	11.8	99.4	11.6

Table 5: The Neg and Pos column indicate % of instances for which all answer choices are predicted as negative or positive. The Incor as Neg, Cor as Pos, and Accurate column indicate % of instances for which all incorrect answers are predicted as negative, the correct answer is predicted as positive, and all answers are predicted accurately as negative or positive. Accurate is the intersection of Incor as Neg and Cor as Pos.

(i.e. MNLI or QNLI) in the commonsense QA task, we performed an analysis on the training data in **NLI-KB** (Huang et al., 2021) and the used CKGs. Following (Mishra et al., 2021), we first compare the vocabulary overlap of ConceptNet, ATOMIC and MNLI (training data) with our evaluation QA datasets. We follow the definition of overlap introduced in (Mishra et al., 2021). Table 6 shows that MNLI has higher vocabulary overlap with all the evaluation datasets than both used CKGs. However, the results for NLI-KB in Table 1 show that the vocabulary overlap is not a key factor for performance as otherwise, NLI-KB fine-tuned with the NLI datasets (before injecting knowledge) should perform better that the other models in the downstream task due to the high lexical similarity.

	Concept	ATOMIC	MNLI
COPA (dev)	50.4	70.0	98.0
COPA (test)	52.1	71.9	86.4
OpenbookQA (dev)	48.4	54.8	92.1
OpenbookQA (test)	48.8	55.2	93.1
SIQA (dev)	37.3	54.6	94.5
CSQA (dev)	59.1	63.2	85.0
SCT (dev)	41.2	57.5	94.5

Table 6: Vocabulary Overlap

SIQA Example	**Question**: After a long grueling semester, Tracy took the final exam and finished their course today. Now they would graduate. Why did Tracy do this? **Answer**: *complete their degree on time*
MNLI	Because I had a deadline. **This entails I had to finish by that time.**
ATOMIC	**Tracy wants finish before time expires.** because Tracy takes the exam.
ConceptNet	pass class causes **graduation.**

Table 7: Alternative answers for SIQA-question.

We also analyze the distance to the sentence embeddings. Our results show that the MNLI entries performed poorly in commonsense knowledge retrieval for SIQA-queries as they are not reasonable answers. In contrast, the sentences generated from ATOMIC and ConceptNet successfully pair the SIQA-questions with reasonable answers. This reveals that, although MNLI has a higher lexical coverage, MNLI does not have suitable examples to match SIQA questions. Thus models fine-tuned with the NLI dataset hardly get any benefit for downstream commonsense reasoning tasks. Tables 7 and 8 present a random sample showing this, where reasonable alternatives are in bold.

CSQA Example	**Question**: If you have leftover cake, where would you put it? **Answer**: *refrigerator*
MNLI	In the waste-paper basket. **This entails in the garbage bin.**
	In the middle of the dinner plate (or is it a base drum?) **This entails in the center of the dinner plate.**
	We always keep it in the hall drawer. **This entails it's always kept in the drawer in the hall.**
ATOMIC	John cuts the cake. as a result, **John wants put the rest of the cake in fridge**
	John places in the oven. but before, John needed to mix the cake ingredients
	John puts in the fridge. but before, John needed to grab it off the table
ConceptNet	**oven** is the position of cake
	refrigerator is the position of moldy leftover
	fridge is the position of leftover

Table 8: Alternative answers for CSQA question.

5 Conclusion

We presented a framework converting KGs into positive/negative question-answer pairs to train a binary classification model, discriminating whether a sentence is reasonable. Extensive experiments show the effectiveness of our approach, while using a reasonably small amount of data. For future work, we will explore how to better select negative cases.

Limitations

The method to select negative examples could be improved, as randomly selecting negative examples for training might lead to identifying most of examples in the evaluation datasets as reasonable. Secondly, we did not explore using other number of candidates in the training set, we always use 2 candidate answers for each question.

Acknowledgments

This work is supported by the Chang Jiang Scholars Program (J2019032).

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.

Prajjwal Bhargava and Vincent Ng. 2022. Commonsense knowledge reasoning and generation with pre-trained language models: A survey. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12317–12325.

Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439.

Jiaoyan Chen, Yuxia Geng, Zhuo Chen, Ian Horrocks, Jeff Z. Pan, and Huajun Chen. 2021. Knowledge-aware Zero-Shot Learning: Survey and Perspective. In *Proceedings of IJCAI*, pages 4366–4373.

Zi-Yi Dou and Nanyun Peng. 2022. Zero-shot commonsense question answering with cloze translation and consistency optimization. In *The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI)*.

Y Geng, J Chen, X Zhuang, Z Chen, J Z Pan, J Li, and H Chen Z Yuan. 2023. Benchmarking knowledge-driven zero-shot learning. *Journal of Web Semantics*.

Deepanway Ghosal, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2022. Two is better than many? binary classification as an effective approach to multi-choice question answering.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations*.

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. 2020. COMET-ATOMIC 2020: On Symbolic and Neural Commonsense Knowledge Graphs.

Hector J. Levesque. 2011. The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*. AAAI.

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.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2019a. K-bert: Enabling language representation with knowledge graph.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13480–13488.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13507–13515.

Chaitanya Malaviya, Chandra Bhagavatula, Antoine Bosselut, and Yejin Choi. 2020. Commonsense knowledge base completion with structural and semantic context. In *Proceedings of AAAI*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. ArXiv:1809.02789 [cs].

Anshuman Mishra, Dhruvesh Patel, Aparna Vijayakumar, Xiang Lorraine Li, Pavan Kapanipathi, and Kartik Talamadupula. 2021. Looking Beyond Sentence-Level Natural Language Inference for Question Answering and Text Summarization. In *Proceedings of ACL*, pages 1322–1336, Online. Association for Computational Linguistics.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. LSDSem 2017 Shared Task: The Story Cloze Test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51, Valencia, Spain. Association for Computational Linguistics.

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.

J. Z. Pan, G. Vetere, J.M. Gomez-Perez, and H. Wu, editors. 2017. *Exploiting Linked Data and Knowledge Graphs for Large Organisations*. Springer.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. ArXiv:2004.09813 [cs].

Julien Romero, Simon Razniewski, Koninika Pal, Jeff Z. Pan, Archit Sakhadeo, and Gerhard Weikum. 2019. Commonsense Properties from Query Logs and Question Answering Forums. In *Proc. of 28th ACM International Conference on Information and Knowledge Management (CIKM 2019)*, pages 1411–1420.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2018. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. 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.

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, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI Conference on Artificial Intelligence*.

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*, AAAI'17, page 4444–4451. AAAI Press.

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.

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.

Jiawei Wang and Hai Zhao. 2022. ArT: All-round thinker for unsupervised commonsense question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1490–1501, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy Liang, and Jure Leskovec. 2022. Deep bidirectional language-knowledge graph pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Appendix

A KGs, Datasets, and Implementation

This section contains more experimental details. In particular, we give details of the used KGs and datasets. We also discuss implementation details.

ConceptNet

ConceptNet (Speer et al., 2017) is a traditional KG that focuses on taxonomic, lexical and physical relations (e.g., *IsA*, *RelatedTo*, *PartOf*). In our experiment, we employed the CN-82K version which is uniformly sampled from a larger set of extracted ConceptNet entity-relations (Li et al., 2016).

ATOMIC

The ATOMIC KG (Sap et al., 2018) focuses on social-interaction knowledge about everyday events, and thus has a higher coverage in the field of commonsense query answering. It consists of 880K knowledge triples across 9 relations (e.g. xNeed, oEffect, xReact). This includes mentions of topics such as causes and effects, personal feelings toward actions or events, and conditional statements. The ATOMIC dataset is collected and validated completely through crowdsourcing.

As seen in Table 2, in comparison to related works: Ma (Ma et al., 2021) and MICO (Su et al., 2022), our methods used much less data from the CKGs (**~5-8x** Ma, **~2-20x** MICO) while still maintaining competitive performance on the evaluation dataset.

A.1 Generation of QA pairs

The QA pairs were generated using the templates in the ATOMIC paper (Hwang et al., 2020), which is compatible with relations in both ConceptNet and ATOMIC. These templates help to convert KG triples into natural sentences, examples shown in Table 9. The head entity and mapped relation phrases are joined as a question. The correct tail entity and a randomly sampled tail from the dataset are used as the positive and negative answers, respectively, for contrastive learning.

A.2 Evaluation Datasets

We evaluate our framework using five downstream QA tasks: COPA, OpenBookQA, SIQA, CSQA, and SCT, which covere a wide range of topics within commonsense reasoning. Accuracy is used as the evaluation metric. All experiments are perform in an unsupervised setting, where our model are not train on the source task.

Choice of Plausible Alternatives (COPA) (Gordon et al., 2012) is a two-choice question-answer dataset designed to evaluate performance in open-domain commonsense causal reasoning. Each entry contains a premise and two possible answers, the task is to select the answers that most likely have a causal relationship with the premise. The dataset consists 500 questions for both debvelopment and test sets.

OpenBookQA (Mihaylov et al., 2018) is inspired from open book exams that assess human understanding in real life. This QA task requires a deeper understanding about both open book facts (e.g., *metals is a heat conductor*) and a broad common knowledge (e.g., *a steal spoon is made of metal*) to answer questions like: *Which of these objects conducts the most heat: A metal spoon, pair of jeans, or cotton made clothing?* It contains 500 multiple-choice science questions for both development and test sets.

SocialIQA (SIQA) (Sap et al., 2019) contains multiple-choice questions with topics concerned with emotional and social interactions in a variety of everyday situations. Each entry comes with a context, a question, and 3 candidate answers. The questions are generated using the `ATOMIC` KG by converting triples into question sentences using predefined templates, and the answers are crowdsourced. The dataset's development split is used as evaluation dataset, containing 1,954 questions.

CommonsenseQA (CSQA) (Talmor et al., 2019) contains questions focused on various commonsense aspects. Each entry contains a question and five candidate answers. The questions are constructed by crowd workers. The answer candidates include distractors comprised of hand-picked ones or nodes from ConceptNet. The development set is used as evaluation set, containing 1,221 questions.

Story Cloze Test (SCT) (Mostafazadeh et al., 2017) is a LSDSem'17 shared task, evaluating story understanding and script learning. Each entry contains a four-sentence story and two possible fifth sentences, where the model has to pick the most suitable ending for the story. The development set is used as the evaluation set, containing 1572 different stories.

Triple	Source	Negative Triple	Generated QA Pairs
(chopstick, AtLocation, table)	ConceptNet	(bread, is created by, flour)	Q: Chopstick located or found at **A: table** B: flour
(PersonX wants to go to the office, oEffect, get dressed up)	ATOMIC	(PersonX leaves the room, xWant, to go somewhere else)	Q: PersonX wants to go to the office, as a result, PersonX will **A: get dressed up** B: to go somewhere else

Table 9: QA pairs generated by KG Triples

A.3 Implementation details

Our experiments are run on a single A100 GPU card. We use RoBERTa-Large as our backbone model. The training batch size is 196, and the maximal sequence length for training is 64. The learning rate is set to 5e-5 for all experiments. For experiments with the margin ranking loss, η is set to 1. The validation set is evaluated by accuracy and used to select a best model for further evaluation. The models are trained for 20 epochs and early stopped when the change of validation loss is within 1%.

B Ablation Studies

We present the full results for the ablation studies discussed in Section 4.3. Table 3 for the backbone models study; Table 4 for the influence of contrastive learning; and Table 5 for accuracy.

C Data Analysis

In the analysis of the distance to sentence embeddings, we treat each entry in the CKG datasets as possible answers and encode them using the SBERT pre-trained model (*all-mpnet-base-v2*) (Reimers and Gurevych, 2019, 2020). Then, the cosine-similarity between the SIQA question and the encoded sentences is calculated to rank their semantic relatedness.

We retrieved the top 3 answers for each source and listed by similarity score at descending order. Table 10 extends the results presented in Section 4.4; Table 11 show the alternative answers from CKG datasets COPA questions.

SIQA Example	**Question**: After a long grueling semester, Tracy took the final exam and finished their course today. Now they would graduate. Why did Tracy do this? **Answer**: *complete their degree on time*
MNLI	Because I had a deadline. **This entails I had to finish by that time.**
	The professors went home feeling that history had been made. **This entails The professors returned home.**
	They got married after his first year of law school.**This entails Their marriage took place after he finished his first year of law school.**
ATOMIC	**Tracy wants finish before time expires.** because Tracy takes the exam
	Tracy wanted to get a degree. as a result Tracy finishes Tracy's test
	Tracy graduates with a degree. but before, Tracy needed get pass with good marks.
ConceptNet	**pass class causes graduation**
	study ends with the event or action graduate
	graduation because **take final exam**

Table 10: Complete results of alternative answers retrieved from MNLI, ATOMIC and ConceptNet for SIQA question. Reasonable alternatives are in bold.

COPA Example	**Question**: The boy wanted to be muscular. As a result, **Answer**: *He lifted weights.*
MNLI	Emboldened, the small boy proceeded. **This entails the small boy felt bolder and continued.**
	Out of shape, fat boy. **This entails the boy was obese.**
	When Sport Resort won the contract for the construction of a new hotel center for 1200 people around the Olympic Sports Arena (built as a reserve for the future, to have it ready in time for the next championships), Gonzo began to push his weight around, because he felt more secure. **This entails when Sport Resort won the contract for the construction of a new hotel Gonzo felt more secure.**
ATOMIC	John wanted to build his physique. as a result **the boy lifts weights**
	The boy starts working out. as a result, the boy wants to gain more muscle
	The boy starts lifting weights. as a result, the boy will build muscle
ConceptNet	**lift could make use of muscle**
	person desires strong body
	build muscle because exercise

Table 11: Alternative answers from CKGs for COPA question.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7

☐ 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?
Section 1

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

B ☑ Did you use or create scientific artifacts?

Section Appendix

☑ B1. Did you cite the creators of artifacts you used?
Section Appendix

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ 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)?
Not applicable. Left blank.

☐ 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?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ 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.
Section Appendix

C ☑ Did you run computational experiments?

Section 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?
Section Appendix

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?
Section Appendix

☐ 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?
Not applicable. Left blank.

☑ 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.)?
Section Appendix

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

Left blank.

☐ 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.?
No response.

☐ 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)?
No response.

☐ 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?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.