CKBP v2: Better Annotation and Reasoning for Commonsense Knowledge Base Population

Anonymous ACL submission

Abstract

Commonsense Knowledge Bases (CSKB) Population, which aims at automatically expanding knowledge in CSKBs with external resources, is an important yet hard task in NLP. Fang et al. (2021a) proposed a CSKB Population (CKBP) framework with an evaluation set CKBP v1. However, CKBP v1 relies on crowdsourced annotations that suffer from a considerable number of mislabeled answers, and the evaluation set lacks alignment with the external knowledge source due to random sampling. In this paper, we introduce CKBP v2, a new high-quality CSKB Population evaluation set that addresses the two aforementioned issues by employing domain experts as annotators and incorporating diversified adversarial samples to make the evaluation data more representative. We show that CKBP v2 serves as a challenging and representative evaluation dataset for the CSKB Population task, while its development set aids in selecting a population model that leads to improved knowledge acquisition for downstream commonsense reasoning. A better population model can also help acquire more informative commonsense knowledge as additional supervision signals for both generative commonsense inference and zero-shot commonsense question answering. Specifically, the question-answering model based on DeBERTa-v3-large (He et al., 2023b) even outperforms powerful large language models in a zero-shot setting, including ChatGPT and GPT-3.5.

1 Introduction

Recently introduced LLMs have shown a remarkable performance on many reasoning benchmarks (Hoffmann et al., 2022; Chowdhery et al., 2022; Bang et al., 2023; Chan et al., 2023), yet there still exists a need to ensure the alignment between the generation of LLMs with external knowledge at the inference time to avoid hallucination and for safer use (Kim et al., 2022a; He et al., 2023a; Peng et al., 2023). The source of

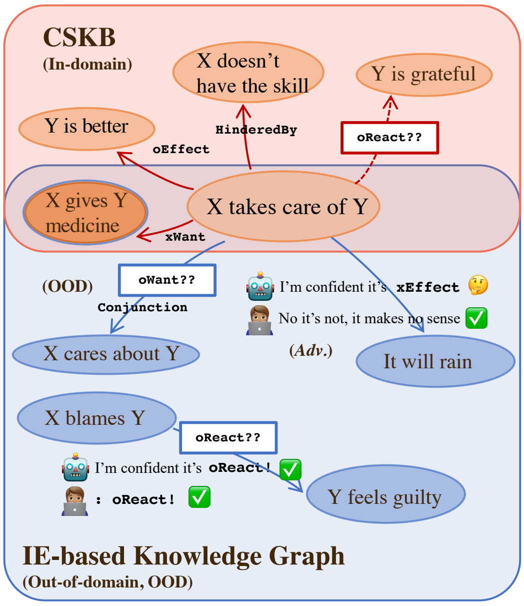

Figure 1: An example of CSKB Population. The coral part indicates the conventional case of CSKB Completion, and the blue part is the population on external knowledge graphs. We include an adversarially constructed sample set in our CKBP v2 by re-annotating the confident predictions by language models.

external knowledge, which can be commonsense, factual, or domain knowledge, should be selected and processed carefully depending on the purpose of generation. However, existing (high-quality) human-annotated knowledge bases are usually far from complete to serve as the source of external knowledge for LLMs.

Regarding commonsense knowledge bases, to extend limited human annotations, CSKB Population (Fang et al., 2021a) stands as a way to acquire missing knowledge, thereby enriching and expanding the existing CSKBs. Unlike CSKB Completion (Li et al., 2016; Saito et al., 2018; Malaviya et al., 2020), which adopts a close-world assumption and only deals with entities and events within CSKBs, the Population task deals with both existing and unseen entities and events, thus requiring a more generalized reasoning ability.

Several works have been conducted on CSKB Population. Fang et al. (2021a) studied a framework that links four CSKBs, ConceptNet (Speer et al., 2017), ATOMIC (Sap et al., 2019a), ATOMIC$^{20}_{20}$ (Hwang et al., 2021), and GLUCOSE (Mostafazadeh et al., 2020), to a large-scale discourse knowledge base, ASER (Zhang et al., 2020, 2022). The resulting knowledge base not only served as the unified source of commonsense knowledge but also was used as the training set to train population models in order to identify missing commonsense knowledge. To evaluate models, the authors created an evaluation set (denoted as CKBP v1), in which they applied fine-grained rules to select candidate commonsense knowledge from ASER and enlisted human annotators to manually annotate these candidates.

However, there are two major limitations in CKBP v1. First, the quality of CKBP v1 is limited. CKBP v1 instances are randomly sampled from the whole population space, resulting in a low recall of plausible commonsense knowledge due to the noise in candidate discourse knowledge. Moreover, as pointed out by Davis (2023), current crowdsourced commonsense benchmarks often contain a substantial fraction of incorrect answers, we also find it true for CKBP v1 after manual inspection. For example, annotators frequently make mistakes on some subtle relations such as `xIntent`, which should describe an *intention* instead of a *consequence*. Second, it's unclear how to leverage populated or expanded commonsense knowledge in CKBP to further improve downstream commonsense reasoning. All previous investigations into CKBP stay within the population task itself without generalizing to actual downstream applications.

Therefore, to address the two limitations, this work presents a more high-quality and adversarially constructed evaluation set by expert annotation, and a comprehensive pipeline for conducting a series of downstream experiments. The aim is to leverage the new CKBP benchmark effectively and facilitate improved utilization for downstream commonsense reasoning tasks.

Leveraging the existing framework, we build CKBP v2 by randomly sampling 2.5k instances from CKBP v1 and adding 2.5k adversarial instances, leading to a total of 5k instances as an evaluation set. These instances are then annotated by experts with substantial expertise in machine commonsense. Then, we present both intrinsic and extrinsic experiments based on CKBP v2. We study the performance of both supervised and semi-supervised task-specific models, together with powerful off-the-shelf language models, such as ChatGPT (OpenAI, 2022) and Vera (Liu et al., 2023), and show that the CKBP v2 evaluation set is still challenging even for advanced language models. Moreover, by employing a CSKB Population model that demonstrates satisfactory performance on CKBP v2, we can enrich existing CSKBs with diverse and novel knowledge that significantly benefits downstream reasoning. We present methodologies and experiments on generative commonsense inference (Bosselut et al., 2019) and zero-shot commonsense question answering (Ma et al., 2021), and show that the acquired commonsense knowledge can be valuable augmented data on the original CSKB and lead to improved downstream performance. In particular, CKBP v2-preferred population model exhibits better alignment than CKBP v1 with advancements in generative commonsense inference.

In summary, our contributions are three-fold: First, We introduce a new evaluation benchmark CKBP v2 for the CSKB Population task, which addresses the quality issues of its predecessor CKBP v1. Second, We launch a pioneer study to use populated commonsense knowledge as additional supervision signals to help downstream commonsense reasoning. Third, We conduct extensive experiments and evaluations with different models on both CKBP v2 itself as well as downstream generative commonsense inference and zero-shot question answering. The results show that CKBP v2 is still a hard task for language models, and the acquired populated knowledge can improve language models' (zero-shot) commonsense reasoning ability on two downstream tasks across six datasets.

2 Related Work

In this section, we discuss 1) CSKBs and their role in the era of LLMs and 2) methods and benchmarks for completing and populating knowledge bases in general.

Commonsense Knowledge Bases. There are many commonsense knowledge bases[1] introduced in the past few years, such as ATOMIC2020 (Hwang et al., 2021), ComFact (Gao et al., 2022), CICERO (Ghosal et al.,

[1]Here, despite the subtle differences between datasets and knowledge bases, we refer to both as knowledge bases

2022), PIQA (Bisk et al., 2020a), Numersense (Lin et al., 2020). Unlike the decades-old knowledge base ConceptNet (Liu and Singh, 2004) that only focuses on taxonomic commonsense, these knowledge bases study a broad range of commonsense, including human-event-centric, contextualized, physical, numerical commonsense.

Along with pure-symbolic CSKBs whose knowledge is obtained from corpora and stored in textual format, there is a stream of research that works on developing neural(-symbolic) CSKBs, which are either knowledge models such as COMET (Bosselut et al., 2019) or symbolic CSKBs built by prompting knowledge from language models, such as ATOMIC^{10X} (West et al., 2022a), SODA (Kim et al., 2022a). Although the approach seems highly scalable and seems promising to build more and larger CSKBs, knowledge from neural(-symbolic) CSKBs remains unreliable (Kim et al., 2022a; He et al., 2023a; Peng et al., 2023) thus often needs to have a robust critic model to filter for good/correct knowledge.

Completing and Populating Knowledge Bases. Regarding conventional knowledge bases like Wordnet (Miller, 1995) and Freebases (Bollacker et al., 2008), tasks involving completion and population have been well-studied as transductive and inductive link prediction problems in the field of graph neural network (Bordes et al., 2013; Yang et al., 2015; Sun et al., 2019; Shang et al., 2019; Fang et al., 2021b). Methods powered by pre-trained language models have also been studied in these tasks thanks to the models' representation power (Yao et al., 2019). In that setting, knowledge instances of the knowledge bases are serialized to a text sequence, which serves as input to LMs such as BERT or RoBERTa.

Specific to CSKB Population task on CKBP v1, Fang et al. (2021a) proposed KGBertSAGE, a combination of KG-BERT (Yao et al., 2019) and GraphSAGE (Hamilton et al., 2017). The model showed higher performance over baselines yet still suffered from the out-of-domain problem. The follow-up work PseudoReasoner (Fang et al., 2022) employs the pseudo-labeling technique to solve that problem. Despite the significant gain in performance, PseudoReasoner is still far from human performance, suggesting that CKBP remains a challenging task in commonsense reasoning.

3 Dataset Construction

In this section, we introduce the task definition, the preparation of the candidate evaluation set, annotation guidelines, and data analysis.

3.1 Task Definition

The task of CKBP (Fang et al., 2021a) is defined as follows. Given $G^C = \{(h, r, t) | h \in H, r \in R, t \in T\}$ (where H, R, T is the set of head events, relations, and tail events), the graph-like knowledge base formed by aligning a union of commonsense knowledge bases C and a much larger discourse knowledge graph G into the same format; the goal of CSKB population task is to learn a scoring function that gives a candidate knowledge triple (h, r, t) higher score if the triple is plausible commonsense. The training process is formulated as triple classification, with ground-truth positive triples from the CSKB C and negative triples randomly sampled from $G^C - C$ with an equal amount. The model is then evaluated on a human-annotated evaluation set E. Here, CKBP v2 serves as the evaluation set.

3.2 Dataset Preparation

We randomly sampled 2.5k instances from CKBP v1 and 2.5k adversarial instances to form CKBP v2. Instances from CKBP v1 are sampled so that the ratio of the number of triples between relations remains unchanged. Meanwhile, the adversarial instances are ones from the candidate knowledge base ASER that the finetuned baseline KG-BERT (Yao et al., 2019) model confidently believes they are plausible, i.e., receives plausibility score $\geq$ 0.9. To ensure the diversity of adversarial instances and hence the evaluation set, we adopt an additional diversity filter using self-BLEU following West et al. (2022a). The triples annotated as negative are considered *hard negatives* as they are what a standard CSKB Population model would favor. Note that we only consider instances of 15 relations other than `general Want/React/Effect`, because most of the triples on the three relations are broken sentences in CKBP v1. We also remove samples of these relations in the training set.

3.3 Annotation Process

Setup We recruited four human experts for the annotation work. The experts are graduate NLP researchers with at least one year of experience working on CSKBs. We randomly divide 5k samples into 4 parts, then for i from 0 to 3, assign the i^{th}

	# Triples	% Plau.	% Unseen
split			
Dev	958	20.46	56.79
Test	4,048	22.06	60.43
instance type			
In-Domain	845	34.56	43.79
Out-of-Domain	1,653	11.92	63.37
Adv.	2,508	23.92	61.12
relation			
xWant	611	22.75	54.01
oWant	239	25.94	58.18
xEffect	603	29.68	55.23
oEffect	172	21.51	58.91
xReact	533	20.64	51.18
oReact	183	13.66	50.70
xAttr	605	23.47	52.91
xIntent	239	16.32	58.40
xNeed	378	25.66	55.37
Causes	236	21.61	55.41
xReason	5	40.0	30.0
isBefore	157	28.03	54.80
isAfter	182	24.73	55.40
HinderedBy	777	12.1	63.17
HasSubEvent	86	26.74	61.04

Table 1: Statistics of CKBP v2. # Triples, % Plausible, and % Unseen, respectively, indicate the number of triples in the subset, the proportion of plausible triples after label finalization, and the proportion of nodes that do not appear in the training set.

and $(i+1 \bmod 4)^{th}$ parts to the i^{th} expert. In this way, two different annotators annotate each triple, and we can fully compare the pairwise agreement between all four annotators. Experts are provided with knowledge triples in the format of (h, r, t), referencing the definition and examples of all relations in Hwang et al. (2021). We ask annotators to judge the plausibility of triples in a three-point Likert scale with corresponding scores: Always/Often (1), Sometimes (0.5), Rarely/Never/Ambiguous/Invalid (0). The final label of an instance is determined as *plausible* if and only if it receives at least one score of 1 and the other score is at least 0.5. For remaining cases, the final label is *implausible*. After finalizing the annotation, we split the evaluation set into development and test sets with a ratio of 1:4 with the preservation of distribution w.r.t labels, relations, and instance types. To estimate human performance, we treat expert annotations as two sets of predictions and compare them to the final labels.

Similar to CKBP v1, we categorize the evaluation set into three groups based on their origin, which are 1) ID: in-domain, whose head and tail events are all from CSKBs, 2) OOD: out-of-domain, which has at least one event outside of CSKBs (equivalent to "CSKB head + ASER tail" and "ASER Edges" in CKBP v1), and 3) *Adv.*: adversarial examples newly introduced in CKBP v2.

Quality Control Although annotators are experts with a clear understanding of the CSKB Population, we acknowledge the ambiguity of CSKB relations and the difficulty in discriminating between them. To control the quality, we provide guidance as a list of scoring criteria. We also carried out a dry run, which asked them to annotate 60 instances covering all relations in order to establish a unified understanding of the problem among participants.

After that, we carry out the main round, where the annotators perform their jobs individually and independently. Throughout the process, we regularly conduct random checks on the samples and engage in discussions with annotators to address any disagreements. We then use the insights gained from these discussions to update and refine our guidance iteratively. After the individual annotation, we facilitated a conflict resolution session to address instances with contrasting scores of 1 and 0. After resolving conflicts, we have the average inter-annotator agreement score IAA as 90.55%.

3.4 Data Analysis

The overall statistics of CKBP v2 are shown in Table 1. It can be easily observed that the new evaluation set has data imbalance issues. However, we do not down-sample the evaluation set to achieve the data balance since the imbalance better reflects the true distribution of plausible and implausible commonsense knowledge in ASER. Given this imbalance, we notice that the AUC scores of examined population models will naturally be high. Also, in the real application of population models, we focus on the precision and recall of the detection for plausible commonsense instances. Thus, in Section 4, along with AUC, we also report the binary F1 scores for each experimented model.

4 Intrinsic Evaluation

4.1 Setup

We examine several models which were previously evaluated on CKBP v1, including zero-shot GPT models (Radford et al., 2019), supervised-learning baselines KG-BERT (Yao et al., 2019) and COMET (Bosselut et al., 2019), and semi-supervised-learning models PseudoReasoner (Fang

Category	Model	AUC				F1			
		all	ID	OOD	*Adv.*	all	ID	OOD	*Adv.*
Zero-shot	GPT2-large	56.47	56.60	58.31	54.22	35.37	47.40	24.06	36.84
	GPT2-XL	56.79	54.47	56.70	54.63	35.22	47.62	23.49	36.65
	GPT3 `text-davinci-003`	61.63	65.93	59.17	59.98	39.44	51.09	28.57	38.20
	ChatGPT `gpt-3.5-turbo`	65.77	70.37	62.56	62.27	45.93	62.59	44.79	26.86
Supervised Learning	KG-BERT (BERT-base)	71.33	84.60	64.47	62.9	45.03	69.27	26.53	41.97
	KG-BERT (RoBERTa-L)	73.70	85.53	67.70	65.60	46.70	69.73	30.73	43.27
	COMET (GPT2-L)	70.00	79.02	66.43	62.62	45.55	61.90	32.14	42.15
	COMET (GPT2-XL)	70.32	79.66	66.53	63.22	45.32	63.34	31.18	40.83
	Vera (T5-xxlarge)	72.45	78.84	68.40	**68.16**	**52.13**	**71.73**	**36.74**	**50.02**
Semi-Supervised	PseudoReasoner BERT-base	71.93	84.23	66.67	63.43	45.47	68.67	30.17	41.77
	PseudoReasoner RoBERTa-L	**74.33**	**85.57**	**69.33**	66.37	46.63	69.70	30.87	43.13
Human		94.1	94.9	91.4	94.5	91.5	94.3	86.9	91.5

Table 2: Main experimental results on CKBP v2. Both AUC and F1 are used as evaluation metrics. The "all" column indicates the overall performance, and ID, OOD, *Adv.* indicate the performance of the In-domain, Out-of-domain, and Adversarial subset. The best results are **boldfaced**, and the second-best ones are underlined.

et al., 2022) with two backbone encoders, BERT-base-uncased (Devlin et al., 2019) and RoBERTa-large (Liu et al., 2019). We use Huggingface[2] Transformers (Wolf et al., 2020) to build our code base. For discriminative models, we set the learning rate as 1e-5, batch size 64/32 for base/large variants, respectively, and the number of training epochs as 1. For generative models (COMET), we use learning rate 1e-5 and batch size 32 to train in 3 epochs. Negative perplexity scores are used as the final prediction scores. For PseudoReasoner, we adopt the best settings in Fang et al. (2022), where we first finetune the KG-BERT model on pseudo-labeling data for one epoch, then from the best checkpoint, we resume the finetuning process on the original training data. Note that the training data and unlabeled data are taken from Fang et al. (2022). We run each baseline three times with different random seeds, then average the result and report in Table 2. For GPT3 (Brown et al., 2020a) and ChatGPT experiments, we use simple prompts asking them to decide whether an assertion is plausible or not.

4.2 Result and Analysis

The results are shown in Table 2. We provide the AUC score and F1 score of all the baselines on the test set in terms of overall performance (all), performance on the subset of ID, OOD, and *Adv.* samples. When calculating F1, for discriminative models, we set the decision threshold as 0.5 (as default), while for generative models, as perplexity serves as the final prediction score, we tune the threshold to obtain the highest F1 score on the development set for each run.

[2] https://huggingface.co/

In the zero-shot setting, the scores increase by the version of GPT. GPT3 `text-davinci-003` gives a significant improvement over GPT2 models, and ChatGPT surpasses its sibling `text-davinci-003` with a similar margin of improvement. Nonetheless, despite the performance improvement from ChatGPT, there is still a clear gap between the zero-shot and (semi-)supervised settings.

In terms of supervised and semi-supervised learning, we observe different scenarios between KG-BERT's performance and COMET's performance, comparing to the result on CKBP v1 reported in Fang et al. (2022). Here, on CKBP v2, KG-BERT outperforms COMET with a significant gap of 3 AUC overall and also outperforms in all subsets of the test set. This shows the importance of including negative (implausible) examples in the training for discriminating commonsense. This also explains why there is no significant improvement of PseudoReasoner over the baseline KG-BERT on this new evaluation set.

4.3 Artifacts Analysis

There is an uprising acknowledgment of "artifacts" (Gururangan et al., 2018; Poliak et al., 2018; Gardner et al., 2021) in a dataset, in other words, spurious correlations or confounding factors between the surface properties of textual instances and their labels, that may incidentally appear in the

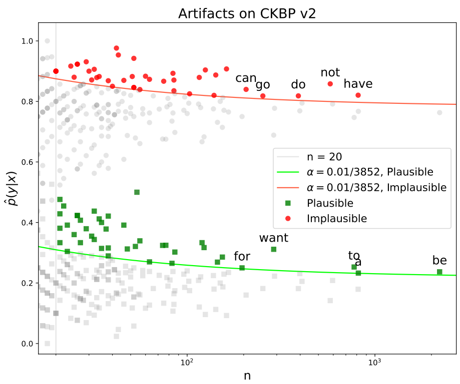

Figure 2: Artifacts statistics of CKBP v2. Colored dots (either square or circle) represent artifacts in the new evaluation set.

annotation process. "Artifacts" may undermine the designated evaluation purpose of the dataset. Thus, it is necessary for us to check if "artifacts" exist in CKBP v2.

We identify artifacts in CKBP v2 by following the previous work Gardner et al. (2021). Particularly, for each word x in the vocab list[3], we compute all quantities appearing in the z-statistic formula

$$z = \frac{\hat{p}(y|x) - p_0}{\sqrt{p_0(1-p_0)/n}}.$$

These include word count n, estimated probability $\hat{p}(y|x)$ as the fraction of the number of target label y in the corresponding n samples over n. After that, we compute the z-statistic and reject or not reject the null hypothesis $\hat{p}(y|x) = p_0$ with a significance level $\alpha = 0.01$ and a conservative Bonferroni correction (Bonferroni, 1936) for all 3852 vocabulary items. Note that the "true" probability $p_0 = p(y|x)$ is taken to be the proportion of samples with label y in the whole evaluation set. Also, we do not consider artifacts with a word count less than 20, as they are not statistically significant.

Figure 2 shows the plot of word count against the estimated probability $\hat{p}(y|x)$ for CKBP v2. The additional green and red curves correspond to the largest value of $\hat{p}(y|x)$ w.r.t n to keep the null hypothesis from being rejected, where y takes value "Plausible" and "Implausible" respectively. This means that any dot above the corresponding curve with a frequency of at least 20 is marked as an artifact. The artifacts with the largest word count are labeled in the plot. Overall, CKBP v2 contains relatively few artifacts (83 artifacts out of 3852 vocabulary items), and the artifacts do not significantly affect the evaluation set quality as their frequencies are not high.

[3]We exclude all relation tokens, as well as special pronoun tokens, namely PersonX, PersonY, PersonZ, PeopleX

5 Extrinsic Evaluation

In this section, we study two downstream applications of CKBP. After acquiring a population model, it act as a scoring function to determine whether a triple from the candidate knowledge base G is plausible or not, thus serving as a source of commonsense knowledge acquisition (Fang et al., 2021b). We leverage the populated knowledge as additional training data for both generative commonsense inference (COMET; Bosselut et al., 2019) and zero-shot commonsense question answering (Ma et al., 2021).

5.1 Generative Commonsense Inference (COMET)

Setup We follow the basic settings as in the original ATOMIC$^{20}_{20}$ paper (Hwang et al., 2021) to generate commonsense tails t given head h and relation r as input. The evaluation dataset is the annotated 5,000 test examples provided by Hwang et al. (2021). We use BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Lavie and Agarwal, 2007), and CIDEr (Vedantam et al., 2015) as the automatic evaluation metrics.

Specifically, we compare the performance of the following training paradigms: 1) Training the model using the official training set of ATOMIC$^{20}_{20}$. 2) Pre-training the model using a comparable amount of CKBP-acquired data, and subsequently fine-tune on ATOMIC$^{20}_{20}$ training set. 3) Training on a mixture of CKBP-acquired data and ATOMIC$^{20}_{20}$ training data.

We filter the CKBP-acquired data using two filters. First, we employ two typical population models, RoBERTa-L (Liu et al., 2019) fine-tuned on CKBP training set and Vera (Liu et al., 2023) to provide a plausibility score for each triple. We set an empirical threshold of 0.8 and selecting triples with plausibility score higher than that as populated commonsense knowledge. For the RoBERTa-L model, we select the best-performed checkpoints based on both CKBP v1 and CKBP v2 to evaluate which evaluation set is better aligned with downstream performance. Second, we utilize a diversity filter defined in G-DAUG (Yang et al., 2020), which is a heuristic favoring diverse n-grams. The diversity filter is applied such that we select the same amount of CKBP-acquired data as the training set

Training Data	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
ATOMIC	41.8	26.6	19.2	14.5	50.0	21.2	66.1
ATOMIC + CKBP$_{\text{RoBERTa-L (V1)}}$	41.9	26.6	18.8	13.8	49.7	21.2	66.2
ATOMIC + CKBP$_{\text{RoBERTa-L (V2)}}$	42.5	26.7	18.8	13.8	50.2	21.4	67.1
ATOMIC + CKBP$_{\text{vera}}$	42.9	27.2	19.4	14.4	50.2	21.4	**67.5**
ATOMIC + CKBP$_{\text{vera (mix)}}$	**43.3**	**27.6**	**19.7**	**14.7**	**50.3**	**21.5**	67.4

Table 3: Performance (%) of GPT2-Large on generative commonsense inference modeling (COMET). ATOMIC stands for ATOMIC$^{20}_{20}$ training set, and CKBP stands for our CKBP data. Subscripts under CKBP indicating the population model to select populated commonsense knowledge. The best performances are **bold-faced**.

of ATOMIC$^{20}_{20}$.

We choose GPT2-Large as our backbone language model. We didn't use GPT2-XL as in Hwang et al. (2021) because the XL version performs relatively poorer than the Large version in terms of most automatic evaluation metrics on the evaluation set of ATOMIC$^{20}_{20}$ despite twice the model size. The learning rate is set as 1e-5, and we train the model for three epochs on both CKBP-acquired data and ATOMIC$^{20}_{20}$ training data.

Results and Analysis The results of generative commonsense inference are presented in Table 3. First, adding CKBP-acquired commonsense knowledge for either pre-training or co-training can yield a general performance improvement in generative commonsense inference. Specifically, the model trained on ATOMIC + CKBP $_{\text{Vera}}$ achieves the best performance and outperforms that only fine-tuned on ATOMIC$^{20}_{20}$ on all automatic evaluation metrics. This indicates that leveraging the abundant unlabeled discourse knowledge from ASER (G), accompanied by appropriate plausibility filtering through the population model, can effectively serve as valuable augmented data to enhance commonsense reasoning. Among the population models, we observe that a better population model, as evaluated by our CKBP v2 evaluation set, corresponds to a higher performance gain in the generative commonsense inference task. This finding highlights the promising potential of developing improved population models, which subsequently contribute to enhanced downstream applications.

Second, the RoBERTa-L model selected by CKBP v2 demonstrates greater efficacy in enhancing generative commonsense inference compared to the model selected by CKBP v1. This finding suggests that CKBP v2 exhibits improved alignment with real-world downstream applications, surpassing its predecessor in terms of practical utility. It's also noteworthy that COMET is an important task that inherently benefits a pile of further downstream tasks that requires commonsense reasoning, including zero-shot commonsense question answering with self-talk (Shwartz et al., 2020) and dynamic graph construction (Bosselut et al., 2021), narrative reasoning (Peng et al., 2022), and dialogue generation (Tu et al., 2022). In this regard, our work exhibits significant potential for generalization to tasks extending beyond the realm of commonsense reasoning.

5.2 Zero-shot Commonsense QA

Setup For the zero-shot commonsense question answering (QA) task, we adopt the task definition and evaluation pipeline proposed by Ma et al. (2021) to evaluate the benefit CKBP v2 brings to extrinsic QA. Several methods have been proposed to tackle this task, including those by Shwartz et al. (2020); Bosselut et al. (2021); Kim et al. (2022b) The most effective pipeline, as proposed by Ma et al. (2021), injects commonsense knowledge into pre-trained language models through fine-tuning on QA pairs synthesized from knowledge in CSKBs. To perform this fine-tuning, the head h and relation r of a (h, r, t) triple are transformed into a question using natural language prompts, while the tail t is used as the correct answer option. Distractors or negative examples are created by randomly sampling tails from triples that do not share common keywords with the head. This fine-tuning process enhances the model's knowledge not only for QA benchmarks constructed from CSKBs, such as SocialIQA (Sap et al., 2019b) derived from ATOMIC, but also improves its ability to answer previously unseen commonsense questions in a more generalized manner.

We adopt the original QA synthesis and model training pipeline by Ma et al. (2021) on the original ATOMIC and the one augmented with populated knowledge from CKBP v2 to ablatively study the sole benefit that knowledge in CKBP v2 brings. Similar with that in COMET experiments, we use the best-performed CKBP model, Vera, to score the whole population space in ASER and select

Model	CSKB	a-NLI	CSQA	PIQA	SIQA	WG	Avg.
Zero-shot Baselines							
Random	-	50.0	20.0	50.0	33.3	50.0	40.7
Majority	-	50.8	20.9	50.5	33.6	50.4	41.2
RoBERTa-L (Liu et al., 2019)	-	65.5	45.0	67.6	47.3	57.5	56.6
DeBERTa-v3-L (He et al., 2023b)	-	59.9	25.4	44.8	47.8	50.3	45.6
Self-talk (Shwartz et al., 2020)	-	-	32.4	70.2	46.2	54.7	-
COMET-DynGen (Bosselut et al., 2021)	ATOMIC	-	-	-	50.1	-	-
SMLM (Banerjee and Baral, 2020)	*	65.3	38.8	-	48.5	-	-
MICO (Su et al., 2022)	ATOMIC	-	44.2	-	56.0	-	-
STL-Adapter (Kim et al., 2022b)	ATOMIC	71.3	66.5	71.1	64.4	60.3	66.7
Backbone: DeBERTa-v3-Large *435M*							
DeBERTa-v3-L (MR) (Ma et al., 2021)	ATM-10X	75.1	71.6	**79.0**	59.7	71.7	71.4
DeBERTa-v3-L (MR) (Ma et al., 2021)	ATOMIC	76.0	67.0	78.0	62.1	76.0	71.8
DeBERTa-v3-L (MR) (Ma et al., 2021)	CKBP (our)	**79.2**	69.6	77.9	64.3	**77.2**	**73.6**
Large Language Models							
GPT-3.5 (`text-davinci-003`)	-	61.8	68.9	67.8	68.0	60.7	65.4
ChatGPT (`gpt-3.5-turbo`)	-	69.3	**74.5**	75.1	**69.5**	62.8	70.2
Supervised Learning & Human Performance							
RoBERTa-L (Supervised)	-	85.6	78.5	79.2	76.6	79.3	79.8
DeBERTa-v3-L (Supervised)	-	89.0	82.1	84.5	80.1	84.1	84.0
Human Performance	-	91.4	88.9	94.9	86.9	94.1	91.2

Table 4: Zero-shot evaluation results (%) on five commonsense question answering benchmarks. The best results are **bold-faced**, and the second-best ones are underlined. The performance of supervised learning and human are for reference only.

the populated knowledge with plausibility scores of over 0.8. Then the same diversity filter as in Section 5.1 is used to downsample the number of populated triples to be comparable with the size of the training set in ATOMIC$^{20}_{20}$. For the QA model, DeBERTa-v3-Large (He et al., 2023b) is used as the backbone, and we train the model using a learning rate of 7e-6 for one epochs on both the CKBP-acquired data and ATOMIC-synthesized data as provided by Ma et al. (2021).

Once trained, we evaluate the model on the validation splits of five commonsense QA benchmarks: Abductive NLI (aNLI; Bhagavatula et al., 2020), CommonsenseQA (CSQA; Talmor et al., 2019), PhysicalIQA (PIQA; Bisk et al., 2020b), SocialIQA (SIQA; Sap et al., 2019b), and WinoGrande (WG; Sakaguchi et al., 2021). Accuracy is used as the evaluation metric. Furthermore, we compare our model not only against existing zero-shot knowledge injection methods (Shwartz et al., 2020; Bosselut et al., 2021; Banerjee and Baral, 2020; Su et al., 2022; Kim et al., 2022b; Ma et al., 2021) but also against large language models such as ChatGPT (OpenAI, 2022) and GPT-3.5 (Brown et al., 2020b).

Results and Analysis The zero-shot commonsense QA results are shown in Table 4. Among all the zero-shot methods, the model trained on CKBP v2 demonstrates the highest performance. It outperforms models trained solely on ATOMIC (with an increase of 2.2%) and ATOMIC10X (West et al., 2022b) (with an increase of 1.8%). Importantly, our method surpasses large language models by an average of 3.4%. This performance gain highlights the significant advantage of our populated commonsense knowledge over both human annotations and distilled knowledge from large language models. Furthermore, we observe that the model trained on CKBP-acquired data shows the most improvement on the aNLI and WinoGrande benchmarks. One potential reason for this is that the populated knowledge in CKBP v1 encompasses a wider range of commonsense knowledge beyond only social commonsense, which benefits tasks involving abductive reasoning (based on narrative) and pronoun coreference resolution.

6 Conclusion

In this paper, we introduce a new CSKB Population benchmark CKBP v2 which addresses two problems of the predecessor CKBP v1. Besides, we conduct a broad range of experiments with different models, including GPT3.5 and ChatGPT, on the new evaluation set. The result shows that the CSKB Population task remains a hard task of commonsense reasoning even for state-of-the-art LLMs, which challenges the community for future research.

Limitations

We observe several limitations of this work. First, CKBP v2 still follows the lemmatized format of events, which may hinder the usage of the resulting population model on knowledge bases other than ASER. Second, the paradigm of CSKB is context-free, which may have difficulty in directly applying to actual downstream tasks. Third, As this paper focuses on proposing a new evaluation set of the CSKB Population, we do not present novel tailored methods for solving this task, leaving it to future research.

Ethical Statements

This work presents CKBP v2, an open-source benchmark for the research community to study the CSKB population problem. The training set is directly adapted from CKBP v1 and ATOMIC$\binom{20}{20}$, GLUCOSE, and ConceptNet, which would have the same ethical issues as in those previous works. Instances in the evaluation set are retrieved from CKBP v1 and ASER, both being open-source with an MIT license. Events in all data instances are anonymized. Thus, the benchmark does not pose any privacy problems about any specific entities (e.g., a person or company). We carried out human expert annotation, where annotators are fairly paid according to the minimum wage requirement of the local government.

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 2020, Online, November 16-20, 2020*, pages 151–162. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020a. PIQA: reasoning about physical commonsense in natural language. 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 7432–7439. AAAI Press.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020b. PIQA: reasoning about physical commonsense in natural language. 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 7432–7439. AAAI Press.

Kurt D. Bollacker, Colin Evans, Praveen K. Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008*, pages 1247–1250. ACM.

C. Bonferroni. 1936. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. 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 2787–2795.

Antoine Bosselut, Ronan Le Bras, and Yejin Choi. 2021. Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering. 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 4923–4931. AAAI Press.

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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2023. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *CoRR*, abs/2304.14827.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Ernest Davis. 2023. Benchmarks for automated commonsense reasoning: A survey. *CoRR*, abs/2302.04752.

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, Quyet V. Do, Hongming Zhang, Yangqiu Song, Ginny Y. Wong, and Simon See. 2022. Pseudoreasoner: Leveraging pseudo labels for commonsense knowledge base population. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3379–3394. 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.

Silin Gao, Jena D. Hwang, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. 2022. Comfact: A benchmark for linking contextual commonsense knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1656–1675. Association for Computational Linguistics.

Matt Gardner, William Merrill, Jesse Dodge, Matthew E. Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. 2021. Competency problems: On finding and removing artifacts in language data. 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 1801–1813. Association for Computational Linguistics.

Deepanway Ghosal, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2022. CICERO: A dataset for contextualized commonsense inference in dialogues. 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 5010–5028. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 107–112. Association for Computational Linguistics.

William L. Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 1024–1034.

Hangfeng He, Hongming Zhang, and Dan Roth. 2023a. Rethinking with retrieval: Faithful large language model inference. *CoRR*, abs/2301.00303.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023b. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *CoRR*, abs/2203.15556.

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.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2022a. SODA: million-scale dialogue distillation with social commonsense contextualization. *CoRR*, abs/2212.10465.

Yu Jin Kim, Beong-woo Kwak, Youngwook Kim, Reinald Kim Amplayo, Seung-won Hwang, and Jinyoung Yeo. 2022b. Modularized transfer learning with multiple knowledge graphs for zero-shot commonsense reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2244–2257. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation, WMT@ACL 2007, Prague, Czech Republic, June 23, 2007*, pages 228–231. 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, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! numersense: Probing numerical commonsense knowledge of pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6862–6868. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226.

Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah A Smith, Yejin Choi, and Hannaneh Hajishirzi. 2023. Vera: A general-purpose plausibility estimation model for commonsense statements. *arXiv preprint arXiv:2305.03695*.

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.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. 2021. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In *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 13507–13515. AAAI Press.

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.

George A. Miller. 1995. Wordnet: A lexical database for english. volume 38, pages 39–41.

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.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *CoRR*, abs/2302.12813.

Xiangyu Peng, Siyan Li, Sarah Wiegreffe, and Mark Riedl. 2022. Inferring the reader: Guiding automated story generation with commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7008–7029, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 180–191. Association for Computational Linguistics.

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

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, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*, pages 141–150. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019a. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *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 3027–3035. AAAI Press.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4462–4472. Association for Computational Linguistics.

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. pages 3060–3067.

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 2020, Online, November 16-20, 2020*, pages 4615–4629. Association for Computational Linguistics.

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.

Ying Su, Zihao Wang, Tianqing Fang, Hongming Zhang, Yangqiu Song, and Tong Zhang. 2022. MICO: A multi-alternative contrastive learning framework for commonsense knowledge representation. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1339–1351. Association for Computational Linguistics.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. MISC: A mixed strategy-aware model integrating COMET for emotional support conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 308–319, Dublin, Ireland. Association for Computational Linguistics.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022a. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4602–4625. Association for Computational Linguistics.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022b. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4602–4625. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. G-daug: Generative data augmentation for commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1008–1025. Association for Computational Linguistics.

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

Hongming Zhang, Xin Liu, Haojie Pan, Haowen Ke, Jiefu Ou, Tianqing Fang, and Yangqiu Song. 2022. ASER: towards large-scale commonsense knowledge acquisition via higher-order selectional preference over eventualities. *Artif. Intell.*, 309:103740.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020. 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.