RiddleSense: Reasoning about Riddle Questions Featuring Linguistic Creativity and Commonsense Knowledge

Bill Yuchen Lin   **Ziyi Wu**   **Yichi Yang**   **Dong-Ho Lee**   **Xiang Ren**
{yuchen.lin,ziyiwu,yichiyan,dongho.lee,xiangren}@usc.edu
Department of Computer Science and Information Sciences Institute,
University of Southern California

Abstract

Question: *I have five fingers but I am not alive. What am I?* Answer: *a glove.*

Answering such a riddle-style question is a challenging cognitive process, in that it requires complex commonsense reasoning abilities, an understanding of figurative language, and counterfactual reasoning skills, which are all important abilities for advanced natural language understanding (NLU). However, there is currently no dataset aiming to test these abilities. In this paper, we present RIDDLE-SENSE[1], a new multiple-choice question answering task, which comes with the first large dataset (5.7k examples) for answering riddle-style commonsense questions. We systematically evaluate a wide range of models over the RIDDLESENSE challenge, and point out that there is a large gap between the best-supervised model and human performance — suggesting intriguing future research in the direction of higher-order commonsense reasoning and linguistic creativity towards building advanced NLU systems.

1 Introduction

> " ***The essence of a riddle is to express true facts under impossible combinations.***"
>
> — *Aristotle*, *Poetics* (350 BCE)

A *riddle* is a puzzling question about concepts in our everyday life. For example, a riddle might ask "*My life can be measured in hours. I serve by being devoured. Thin, I am quick. Fat, I am slow. Wind is my foe. What am I?*" The correct answer "*candle*," is reached by considering a collection of *commonsense knowledge*: a candle can be lit and burns for a few hours; a candle's life depends upon its diameter; wind can extinguish candles, etc.

It is believed that the *riddle* is one of the earliest forms of oral literature, which can be seen as
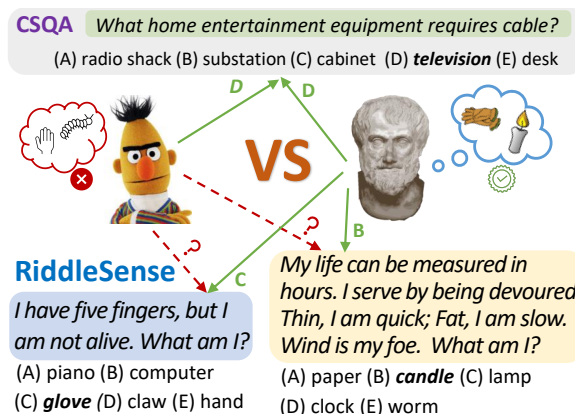

Figure 1: The top example is a trivial commonsense question from the CommonsenseQA (Talmor et al., 2019) dataset. The two bottom examples are from our proposed RIDDLESENSE challenge. The right-bottom question is a descriptive riddle that implies multiple commonsense facts about *candle*, and it needs understanding of figurative language such as metaphor; The left-bottom one additionally needs counterfactual reasoning ability to address the *'but-no'* cues. These riddle-style commonsense questions require NLU systems to have higher-order reasoning skills with the understanding of creative language use.

a formulation of thoughts about common sense, a mode of association between everyday concepts, and a metaphor as higher-order use of natural language (Hirsch, 2014). Aristotle stated in his *Rhetoric* (335-330 BCE) that good riddles generally provide satisfactory metaphors for rethinking common concepts in our daily life. He also pointed out in the *Poetics* (350 BCE): "the essence of a riddle is to express true facts under impossible combinations," which suggests that solving riddles is a nontrivial reasoning task.

Answering riddles is indeed a challenging cognitive process as it requires *complex* commonsense reasoning skills. A riddle can describe *multiple pieces* of commonsense knowledge with *figura-*

tive devices such as metaphor and personification (e.g., "wind is my <u>foe</u> → *extinguish*"). Moreover, *counterfactual thinking* is also necessary for answering many riddles such as "*what can you hold in your left hand <u>but not</u> in your right hand?* → *your right elbow.*" These riddles with *'but-no'* cues require that models use counterfactual reasoning ability to consider possible solutions for situations or objects that are seemingly impossible at face value. This *reporting bias* (Gordon and Van Durme, 2013) makes riddles a more difficult type of commonsense question for pretrained language models to learn and reason. In contrast, *superficial* commonsense questions such as "*What home entertainment equipment requires cable?*" in CommonsenseQA (Talmor et al., 2019) are more straightforward and explicitly stated. We illustrate this comparison in Figure 1.

In this paper, we introduce the RIDDLESENSE challenge to study the task of answering riddle-style commonsense questions[2] requiring *creativity*, *counterfactual thinking* and *complex commonsense reasoning*. RIDDLESENSE is presented as a *multiple-choice question answering* task where a model selects one of five answer choices to a given riddle question as its predicted answer, as shown in Fig. 1. We construct the dataset by first crawling from several free websites featuring large collections of human-written riddles and then aggregating, verifying, and correcting these examples using a combination of human rating and NLP tools to create a dataset consisting of 5.7k high-quality examples. Finally, we use *Amazon Mechanical Turk* to crowdsource quality distractors to create a challenging benchmark. We show that our riddle questions are more challenging than CommonsenseQA by analyzing graph-based statistics over ConceptNet (Speer et al., 2017), a large knowledge graph for common sense reasoning.

Recent studies have demonstrated that fine-tuning large pretrained language models, such as BERT (Devlin et al., 2019a), RoBERTa, and AL-BERT (Lan et al., 2020), can achieve strong results on current commonsense reasoning benchmarks. Developed on top of these language models, graph-based language reasoning models such as KagNet (Lin et al., 2019) and MHGRN (Feng et al., 2020) show superior performance. Most recently, UnifiedQA (Khashabi et al., 2020) pro-

poses to unify different QA tasks and train a text-to-text model for learning from all of them, which achieves state-of-the-art performance on many commonsense benchmarks.

To provide a comprehensive benchmarking analysis, we systematically compare the above methods. Our experiments reveal that while humans achieve 91.33% accuracy on RIDDLESENSE, the best language models can only achieve 68.80% accuracy, suggesting that there is still much room for improvement in the field of solutions to complex commonsense reasoning questions with language models. We believe the proposed RIDDLESENSE challenge suggests productive future directions for machine commonsense reasoning as well as the understanding of higher-order and creative use of natural language.

2   Construction of RIDDLESENSE

In this section, we first present our pipeline for collecting the RIDDLESENSE dataset, including the details of data cleaning. We introduce how we design a crowd-sourcing protocol for annotating quality distractors to turn riddle-solving into a multiple-choice question answering task.

2.1   Riddle Crawling and Cleaning

We write web crawlers for collecting a large number (approximately 10,000) of riddles and their answers from public riddle websites, such as *brainzilla.com*, *riddlewot.com*, etc. As the crawled data contain much noise such as inconsistent answer format and misspelled words, we process riddles through careful data cleaning as well as human verification. First, we use an open-source tool for detecting typos[3] and then refine the sentences. Then we continuously sample (riddle, answer) pairs and recognize errors, for which we iteratively improve our program with a set of conditions to filter out noisy examples that are not readable or have ambiguous answers. Also, we merge the riddles from different sources while removing duplicate riddle questions with similar answers. For detecting duplicate riddles with minor word changes, we use SentenceBERT (Reimers and Gurevych, 2019) to find clusters with high cosine similarities.

---

[2]We use "riddle" and "riddle-style commonsense question" interchangeably in this paper.

[3]`github.com/phatpiglet/autocorrect`

2.2 Distractor Collection from AMT

We consider a multi-choice question answering format rather than the open-ended format, as it is easier to meaningfully compare the performance of different models in a more controlled manner — there is a limited range of of options. For such a dataset, given a riddle-style question and 5 answer options, the model should select the best one as the predicted answer. This format offers a straightforward and fair evaluation metric – *accuracy*, which is the metric adopted by many popular commonsense reasoning benchmarks such as CommonsenseQA, ARC (Clark et al., 2018), and OpenbookQA (Mihaylov et al., 2018).

High-quality distractors are essential for multiple-choice question answering tasks as they can ensure that the dataset is both *clean* and *challenging* — the distractors are neither too similar nor too distant from the correct answer. We thus design a protocol to collect quality distractors from human annotators via *Amazon Mechanical Turk*[4] based on a pool of candidate distractors.

Candidate Distractor Pool We use Q to denote the concepts that are mentioned in the question, and a to denote the concept in the answer[5]. We then first get all two-hop neighbors in the ConceptNet of a and one-hop neighbors of each $c \in Q$ respectively:

$$A = \{x | (x, r_i, y), (y, r_j, a)\}$$
$$B = \{x | (x, r_k, c), \forall c \in Q\}$$
$$D = A \cap B,$$

where $r_{i/j/k}$ is a binary relation in the ConceptNet such as `HasProperty`. The final intersection, D, is thus the pool of distractor candidates. We further use *WordNet* (Miller, 1992) to filter out concepts that have either too low or too high *Wu-Palmer* similarity[6]. We argue that such sampled distractors are semantically relevant to both questions and answers, and are also closer to answers in the WordNet taxonomy. Thus, they are more likely to serve as ideal distractors in a multiple-choice question answering task.

AMT Crowd-sourcing We design a three-stage annotation protocol:

---

[4] https://www.mturk.com/

[5] If there are multiple concepts, we pick the one with the least network degrees as they tend to be more important.

[6] We use 0.5 as a threshold which is effective as expected.

	CSQA	RS
# All Examples	12,102	5,733
# Train Examples	9,741	3,510
# Validation Examples	1,221	1,021
# Test Examples	1,140	1,202
Average Question Length	15.06	24.04
% Long Qs (>20 tokens)	16.5%	47.3%
Distinct Question Words	6,822	7,110
Distinct Choice Words	7,044	9,912
Avg PLL of Qs	-34.41	-53.98
QA-NLI Conflict	12.7%	39.6%
QA-NLI Neutral	71.6%	44.9%
QA-NLI Entailment	15.7%	15.5%

Table 1: Key statistics of the RIDDLESENSE dataset (v1.0) vs the CommonsenseQA (CSQA) dataset.

- S1) **Sanity Check**. We show a question and 3 choices where only 1 choice is correct and the other 2 are randomly sampled concepts from the full vocabulary of ConceptNet. Only when the workers pass this sanity check, their following annotations will be considered, so we can avoid noise from random workers.

- S2) **Candidate Selection**. As it is difficult to control and verify the quality of distractors from crowd workers, we first sample concepts from ConceptNet, which are relevant to both question concepts and answer concepts, forming a set of candidate distractors D for annotators to choose from. Workers are required to select at least 5 concepts that they think are good distractors to the question. There are at least 3 different workers for each question and we take the candidates which are selected by at least two different workers to make sure the selected distractors are indeed meaningful.

- S3) **Open Distractor Collection**. We also ask *master workers* on AMT to write at least one more distractor based on the question context. This stage is important because sometimes the candidate pool contains fewer candidates of good quality and the human-written distractors are usually better than the ones in the candidate pool. We thus give extra bonus credits to encourage annotators to write more quality distractors.

1506

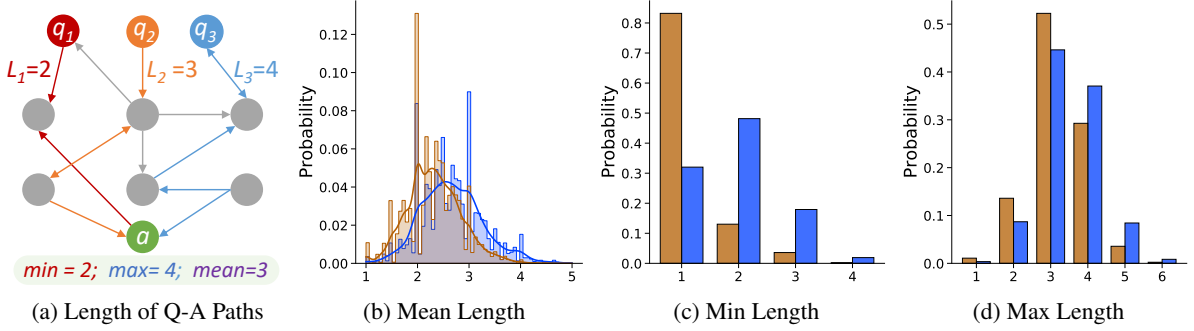(a) Length of Q-A Paths (b) Mean Length (c) Min Length (d) Max Length

Figure 2: The Q-A paths serve as estimation of underlying reasoning chains. Fig. (a) illustrates how to compute mean/min/max of the Q-A paths: $\{q_1, q_2, q_3\}$ are three concepts mentioned in the question, and a is the answer concept. L_k is the length of the shortest path between q_k and a over ConceptNet; *min/max/mean* are computed over $\{L_1, L_2, L_3\}$ as three aspects to measure the overall difficulty. Fig. (b), (c), and (d) show that generally RIDDLESENSE has a longer question-answer path than CommonsenseQA, thus being harder to reason.

Algorithm 1: Get statistics of QA paths.

Input: Knowledge graph $KG = (V, E)$,
　　　　riddle question Q, riddle answer A
Output: $minPathLength$,
　　　　　$maxPathLength$,
　　　　　$meanPathLength$
1　$QC \leftarrow \texttt{extractConcept}(Q)$
2　$AC \leftarrow \texttt{extractConcept}(A)$
3　$ac \leftarrow v \in AC$ with smallest $\texttt{deg}(G, v)$
4　$l \leftarrow []$
5　**foreach** $qc \in QC$ **do**
6　　$path \leftarrow \texttt{shortestPathLen}(KG, qc, ac)$
7　　**if** $path \neq None$ **then**
8　　　$l.\texttt{append}(path)$
9　$minPathLength \leftarrow \min(l)$
10　$maxPathLength \leftarrow \max(l)$
11　$meanPathLength \leftarrow \texttt{mean}(l)$

3　Data Analysis of RIDDLESENSE

In this section, we first report the key statistics of the proposed RIDDLESENSE dataset, then we compare it to CommonsenseQA (Talmor et al., 2019) from two major angles: the distribution of the lengths of Q-A paths and the types of reasoning chains, which serve as an effective proxy to analyze the differences between the two datasets.

3.1　Key Statistics

Table 1 presents the key statistics of RIDDLE-SENSE (RS) and the comparisons with CommonsenseQA (CSQA) which is the most similar benchmark to ours. Although the size of RS is smaller than CSQA, we argue that RS is complementary to the CSQA dataset and introduces novel challenges for the commonsense reasoning community. As they share the same format, we can test different methods by training on either CSQA-only, RS-only, or the concatenation of CSQA and RS, as we show later in Section 4.

Moreover, there is a greater number of long questions (i.e., containing more than 20 words) in RS than in CSQA. Additionally, we find that RS questions have a lower normalized pseudo-likelihood (PLL) (Salazar et al., 2020), a proxy of estimating sentence probability, suggesting that RS questions are more puzzling (i.e., the words are less frequently co-occurring). We also use a RoBERTa model fine-tuned on MNLI (Williams et al., 2018) to perform natural language inference between CSQA/RS questions and their answers. There is a much greater proportion of questions in RS that have *conflicting* relations with their correct answers than compared to CSQA. This is indicative of RS's complexity due to the *self-contradictory* and *perplexing* nature of riddles.

Interestingly, we also find that although there are about twice as many examples in CSQA as RS, there are more distinct words in the questions and answer choices of RS than CSQA, suggesting that RS covers more diverse topics than CSQA.

3.2　Distribution of the Lengths of Q-A Paths

Our main intuition is that the shortest paths between question concepts and the answer concepts can approximate the *underlying reasoning chains*, which are hidden and difficult to label. To understand the difference between CSQA and RS in

CommonsenseQA (CSQA)			
1-hop (14.0%)	**2-hop (34.4%)**	**3-hop (41.5%)**	**4-hop (9.5%)**
AtLoc (4.8%)	Related-Related (8.3%)	Related-Related-Related (4.1%)	Related $\times$ 4 (0.4%)
Related (3.4%)	Related-AtLoc (4.5%)	Related-Related-AtLoc (2.7%)	Related $\times$ 3 -AtLoc (0.3%)
Causes (1.1%)	Related-Antonym (1.8%)	Related-AtLoc^{-1}-AtLoc (1.4%)	Related-Related-AtLoc^{-1}-AtLoc (0.3%)
Antonym (0.9%)	Related-IsA^{-1} (1.3%)	Related-Related-Antonym (1.3%)	Related $\times$ 3 -Antonym (0.2%)
Capableof (0.8%)	Related-AtLoc^{-1} (0.9%)	Related-Related-CapableOf (1.3%)	Related$\times$2-SubEvent^{-1}-Cause (0.1%)
...	...	...	...
$\rho = \frac{3.4}{4.8} = 0.7$	$\rho = \frac{8.3}{4.5} = 1.8$	$\rho = \frac{4.1}{2.7} = 1.5$	$\rho = \frac{0.4}{0.3} = 1.3$
RiddleSense (RS)			
1-hop (4.6%)	**2-hop (31.6%)**	**3-hop (47.8%)**	**4-hop (14.0%)**
Related (3.1%)	Related-Related (13.1%)	Related-Related-Related (10.6%)	Related $\times$ 4 (1.8%)
Antonym (0.4%)	Related-Antonym (2.1%)	Related-Related-IsA^{-1} (2.6%)	Antonym-Related$\times$3 (0.4%)
IsA^{-1} (0.3%)	Related-IsA^{-1} (2.0%)	Related-Related-Antonym (1.6%)	Related$\times$3 -IsA^{-1} (0.3%)
PartOf (0.1%)	Related-AtLoc^{-1} (1.3%)	Related-Antonym-Related (1.5%)	Related$\times$2-IsA^{-1}-Related (0.3%)
AtLoc^{-1} (0.1%)	Antonym-Related (0.8%)	Antonym-Related-Related (1.5%)	Related$\times$2-Antonym-Related (0.3%)
...	...	...	...
$\rho = \frac{3.1}{0.4} = 7.8$	$\rho = \frac{13.1}{2.1} = 6.2$	$\rho = \frac{10.6}{2.6} = 4.1$	$\rho = \frac{1.8}{0.4} = 4.5$

Table 2: The top-5 most frequent types of reasoning chains in CSQA and RS datasets, grouped by their length $k = \{1, 2, 3, 4\}$. The implicit-ratio ρ is defined as the ratio of the implicit reasoning types (i.e., Related$\times k$) over the most frequent types with at least one explicit relation (e.g., AtLoc) of the same length k.

terms of their reasoning chains, we use *Q-A paths* over ConceptNet as a proxy. For a riddle question, a set of *Q-A path lengths* are the lengths of the shortest paths between every question concept and the answer concept, i.e., *shortestPathLen(KG, qc, ac)* in Alg. 1. For a question-answer pair, we first extract the concepts mentioned in the question and the answer respectively (*extractConcept()* in Algorithm 1), following the steps of Lin et al. (2019) and Feng et al. (2020). If there are three question concepts $\{q_1, q_2, q_3\}$ and an answer concept a, we denote their shortest path lengths as $\{L_1, L_2, L_3\}$. Finally, we compute the min/max/mean over them for a comprehensive understanding of the approximated difficulty of this riddle — a greater value indicates a more challenging example.

As shown in Figure 2 (b), we can see that RS has longer Q-A paths as underlying reasoning chains. In addition, we can see that RS generally has longer chains, particularly the min of CSQA is 1-hop for more than 80% of examples. On the other hand, only about 30% of RS examples have 1-hop minimum Q-A paths, while about 50% of the examples have 2-hop min Q-A paths. The distribution over the maximum in Figure 2 (d) also shows that RS tends to have longer maximum paths than CSQA. We also show the percentage of all Q-A paths of different length as part of Table 2, and we can see that RS has longer paths in general (e.g., CSQA = 14.0% vs. RS = 4.6% in 1-hop).

3.3 Relational Types of Reasoning Paths

In addition to the analysis on path length, we also show that the relation types of Q-A paths for RS and CSQA have clear differences, as shown in Table 2. The types of reasoning chains in RS rely more on a special relation in ConceptNet — Related, which is relatively more implicit and can not be grounded to a specific, explicit relation such as AtLoc (e.g., <*wind*, Related, *air*> vs. <*lamp*, AtLoc, *table*>). The most frequent relation between question concepts and answer concepts in CSQA is the AtLoc relation (4.8%), however, it is Related (3.1%) in RS. We define *implicit-ratio* for k-hop paths, $\rho_k = \frac{\%(\text{Related}\times k)}{\%(E_k)}$, where E_k is the most frequent type of chains with at least one explicit relation of length k. In RS, ρ_k is around $4.1 \sim 7.8$, while it is about $0.7 \sim 1.8$ for CSQA. Thus, we conclude that the dominant reasoning chains in RS are much more implicit, and consequently RS is more challenging to reason with using commonsense knowledge resources like ConceptNet.

4   Experiments

We first introduce three types of popular baseline methods for commonsense reasoning (Section 4.1), then we present our main experimental results with analysis (Section 4.2), and finally show case studies for error analysis (Section 4.3).

(1) Fine-Tuning BERT/RoBERTa/ALBERT, etc.

(2) Symbolic KG-based Language Reasoning
KagNet (Lin et al. 2019) & MHGRN (Feng et al. 2020)

(3) Fine-Tuning T5 with a text-to-text task format.
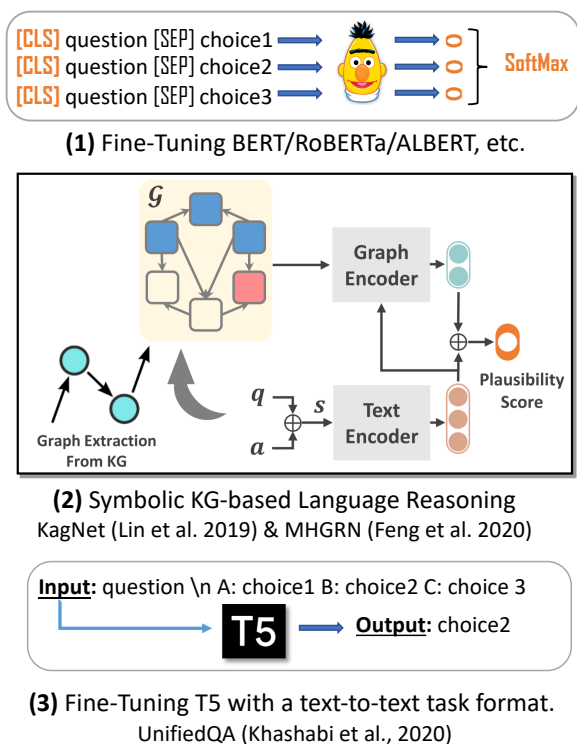UnifiedQA (Khashabi et al., 2020)

Figure 3: Three types of baseline methods: 1) fine-tuning pre-trained LMs, 2) incorporating graph-based reasoner, 3) fine-tuning a unified text-to-text LM.

4.1 Baseline Methods

Given a riddle question q, there are 5 different choices $\{c_1, \ldots, c_5\}$, where only one of them is the correct choice and the others are distractors. The model needs to rank all choices and select the best one as the final answer. There are three major types of models for commonsense reasoning tasks in this format: 1) fine-tuning pretrained language models, 2) incorporating relevant knowledge graphs for reasoning, 3) fine-tuning a unified text-to-text QA model, as shown in Figure 3.

Fine-tuning Pre-trained LMs As we seek to investigate how well current NLU models can perform in higher-order commonsense reasoning, we first experiment with a typical set of large pretrained language models such as BERT (Devlin et al., 2019b), RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2020). We concatenate the question with each choice, using `[SEP]` as the separator, thus forming a *statement*. Then, we fine-tune any pretrained LMs like BERT to use their `[CLS]` token embeddings to predict a score for each statement. Then, a set of five scores about an example will be fed to SoftMax to optimize for maximizing the score of the correct choice.

LMs + Graph Reasoning Modules Kag-Net (Lin et al., 2019) and MHGRN (Feng et al., 2020) are two typical graph-based language reasoning models. They both extract a *schema graph* from ConceptNet, i.e., a subgraph of ConceptNet consisting of Q-A paths in Figure 2, by incorporating them with a graph encoding module. They finally fuse the external commonsense knowledge with a text encoder (e.g., a pretrained LM). Kag-Net uses heuristics to prune irrelevant paths and then encode them with path-based LSTM and hierarchical attention to select the most important paths for improving commonsense reasoning. In contrast, the recent MHGRN explicitly encodes multi-hop paths at scale using graph networks with relational attention, improving efficiency and performance over KagNet and other models. A unique merit of such graph-based models is their *interpretibility* due to the neural attention over the symbolic structures of KGs.

Fine-Tuning a Text-to-Text QA Model UnifiedQA (Khashabi et al., 2020), the state-of-the-art multiple-choice QA model, simply concatenates the question with all answer candidates as a single input sequence to a T5 (Raffel et al., 2020) model for learning to generate the correct choice as extracting a span from the input. Apart from the multiple-choice QA format, it is also trained with other QA task formats so that it can benefit from many other QA datasets (including CSQA) via sharing the model parameters.

Human Evaluation We invite three native English speakers who study computer science to solve 100 riddle examples sampled from the test set. They achieved an average accuracy of 91.3%.

4.2 Results and Analysis

We show the main results of the experiments in Table 3. There are 3 settings according to the different training data options: 1) the training data of CSQA, 2) the training data of RS, and 3) the concatenation of both RS and CSQA, while all experiments are validated over the dev set of RS. However, as the public UnifiedQA checkpoints were already trained on CSQA (together with many other QA datasets), we directly use them for inference over RS in the first setting (i.e., "Train=CSQA"). This also suggests that the performance of UnifiedQA models in 2nd setting should be better than others although they all are fine-tuned on RS's training data only.

Models ↓ Training Data →	Train = CSQA		Train = RiddleSense		Train = RS+CSQA	
RiddleSense-Split →	Dev	Test	Dev	Test	Dev	Test
Random Guess	20.0	20.0	20.0	20.0	20.0	20.0
BERT-Base (Devlin et al., 2019a)	33.59	34.61	54.16	42.43	56.22	47.67
BERT-Large (Devlin et al., 2019a)	36.14	39.10	55.24	45.09	57.69	54.91
RoBERTa-Large (Liu et al., 2019)	43.68	47.42	60.72	52.58	66.11	59.82
ALBERT-XXL (Lan et al., 2020)	**51.03**	**51.00**	<u>66.99</u>	<u>60.65</u>	**71.50**	<u>67.30</u>
KagNet (RB-L) (Lin et al., 2019)	42.66	48.24	61.77	53.72	66.55	59.72
MHGRN (RB-L) (Feng et al., 2020)	46.83	49.65	63.27	54.49	66.90	63.73
MHGRN (AB-XXL) (Feng et al., 2020)	<u>50.89</u>	50.21	66.27	59.93	<u>70.81</u>	66.81
UnifiedQA (T5-L) (Khashabi et al., 2020)	28.50	37.27	56.21	56.40	58.17	56.57
UnifiedQA (T5-3B) (Khashabi et al., 2020)	37.32	<u>50.25</u>	**67.38**	**66.06**	68.26	**68.80**
Human Performance	-	91.33	-	91.33	-	91.33

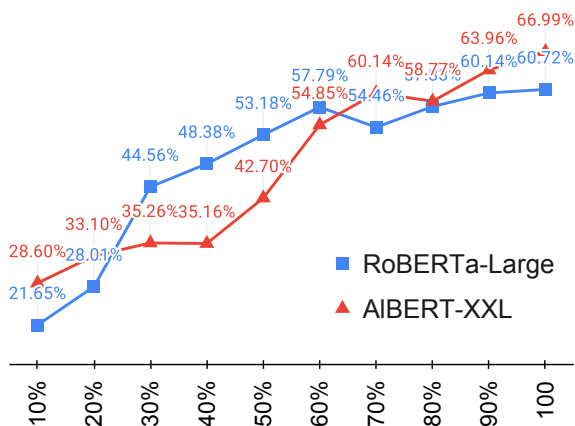Table 3: Benchmark performance over the dev and test set of RIDDLESENSE (v1.0).

Figure 4: The curve of dev accuracy using different percentage of the RS-training data, respectively for RoBERTa-Large and ALBERT-XXL.

We can see that larger pretrained language understanding models always gain better performance, ranging from BERT-base to Albert-XXL, which gets the best performance in this group of baselines (67.30%). This matches their performance comparisions on CSQA and other benchmark datasets as well, suggesting that a better pretrained language model can be also identified by RIDDLESENSE as well. Interestingly, we find that ALBERT-XXL is so powerful that it can generalize from training on CSQA only but achieve comparable results with RoBERTa-Large that is trained over RS (i.e., 51.0% vs. 52.6%). However, if we look at the curve of dev accuracy when using different percentage of the RS-train data (setting 2) in Figure 4, we can see that RoBERTa-Large can generally outperform ALBERTA-XXL when using less than 60% data for fine-tuning.

Moreover, we find that the KG-enhanced models, KagNet and MHGRN, using RoBERTa-Large (RB-L) as the encoder, perform better than vanilla RB-L. Although the Q-A paths over Concept-Net have more implicit paths (e.g., $\text{Related} \times k$), some paths can still be beneficial. For example,

$$wind \xleftarrow{\text{Related}} blow \xleftarrow{\text{Related}} candle,$$

can still help reason about the riddle *"... Wind is my foe. What am I?"* to the answer "candle."

The fusion of ConceptNet also improves in the situation when only training with CSQA data using RoBERTa-Large. However, the improvement of KagNet is negative, which is unexpected. We conjecture that this is because the extracted subgraphs from the ConceptNet does not guarantee the reasoning path from question concepts to answer concepts, while the training phase *forces* models to learn to reason over those graphs, yielding a possibly *harmful* impact. Additionally, we find that MHGRN with ALBERT-XXL also results in a worse performance, unlike using RoBERTa-Large. We believe this may be related to the specific design of ALBERT, which reuses model parameters for multiple layers, and thus it could be a problem when fused with another learnable module (e.g., a graph network in MHGRN).

Fine-tuning UnifiedQA with T5-3B achieves the best performance, which is also the case for CSQA in their leaderboard. This is expected for two reasons: 1) UnifiedQA has been trained over multiple other QA datasets, which increases its

generalization ability, 2) UnifiedQA considers all choices together at a time and thus can better compare different choices with self-attention mechanism of Transformer (Vaswani et al., 2017).

4.3 Error Analysis and Future Directions

We show a few examples that are mistakenly predicted by the UnifiedQA-3B model in Figure 5. From these concrete cases, we can see that even the best model cannot solve riddles that can be trivial to humans, especially when there are metaphors and/or counterfactual situations. We argue that future research should aim to address the creative use of language in commonsense reasoning and general understanding of language, as creativity is a critical feature of natural language. We list several promising directions as follows.

First of all, we should *mine (semi-)structured knowledge of metaphors*, so that concepts can connect via metaphorical links (e.g., "tail" → "thread"). Second, to prevent false inferences, we need *more complete, precise commonsense knowledge of concepts*. For example, in Figure 5, a model should know a chair only has exactly *four* legs instead of *hundreds* (Lin et al., 2020a); ink can be black or red, but it won't change over time. However, current KGs only have (leg, PartOf, chair) and (ink, HasProperty, black/red). In addition, the reasoning methods should incorporate more *symbolic logic rules*, so that the multi-hop conditions and counterfactual "but-no" negations will be handled better. Finally, we think the graph-augmented methods should be improved to *compare multiple options* in a schema graph, e.g., QA-GNN (Yasunaga et al., 2021). Both KAGNET and MHGRN consider only a single option at a time which prevents them from effectively reasoning about the subtle differences between options.

5 Related Work

Benchmarking Machine Common Sense

The prior works on building commonsense reasoning benchmarks touch different aspects of commonsense reasoning: SWAG (Zellers et al., 2018), HellaSWAG (Zellers et al., 2019), CO-DAH (Chen et al., 2019), aNLI (Bhagavatula et al., 2020) for situation-based reasoning; Physical IQA (Bisk et al., 2020) on physical knowledge; Social IQA (Sap et al., 2019) on social psychology knowledge; LocatedNearRE (Xu et al., 2018) on mining spatial commonsense

knowledge; DoQ (Elazar et al., 2019) and NumerSense (Lin et al., 2020a) on numerical common sense; CommonGen (Lin et al., 2020b) for generative commonsense reasoning, and many others; OpenCSR (Lin et al., 2021) and ProtoQA (Boratko et al., 2020) aim to test commonsense reasoning ability in an open-ended setting.

CommonsenseQA (Talmor et al., 2019) has the same format as our proposed RIDDLESENSE, and both target general commonsense knowledge via multiple-choice question answering. However, CSQA focuses more on straightforward questions where the description of the answer concept is easy to understand and retrieval over ConceptNet, while RS makes use of riddle questions to test higher-order commonsense reasoning ability. More detailed comparisions between them are in Section 3, which shows that the unique challenges of the RiddleSense on multiple dimensions.

Commonsense Reasoning Methods

Our experiments cover three major types of commonsense reasoning methods that are popular in many benchmarks: fine-tuning pretrained LMs (Devlin et al., 2019a; Liu et al., 2019; Lan et al., 2020), graph-based reasoning with external KGs (Lin et al., 2019; Feng et al., 2020), and fine-tuning unified text-to-text QA models (Khashabi et al., 2020). Apart from ConceptNet, There are also some methods (Lv et al., 2020; Xu et al., 2020) using additional knowledge resources such as Wikipedia and Wiktionary. A few recent methods also aim to generate relevant triples via language generation models so that the context graph is more beneficial for reasoning (Wang et al., 2020; Yan et al., 2020). Our experiments in this paper aim to compare the most typical and popular methods which have open-source implementations, which we believe are beneficial for understanding the limitation of these methods in higher-order commonsense reasoning — RIDDLESENSE.

Computational Creativity and NLP

Creativity has been seen as a central property of the human use of natural language (McDonald and Busa, 1994). Text should not be always taken at face value, however, higher-order use of language and figurative devices such as metaphor can communicate richer meanings and needs deeper reading and more complicated reasoning skills (Veale, 2011). Recent works on processing language with creative use focus on metaphor detection (Gao

Riddle Questions	Choices (√=truth; ×=model's choice)	Explanation
I am black when you buy me, red when you use me. When I turn white, you know it's time to throw me away. What am I?	(A) **charcoal (√)** (B) rose flower (C) **ink (×)** (D) fruit (E) shoe	**Describing multiple conditions of a common object.** Only charcoal applies to all.
I have a long tail that I let fly. Every time I go through a gap, I leave a bit of my tail in the trap. What am I?	(A) monkey (B) basketball (C) **fishing pole (×)** (D) comet (E) **needle (√)**	**Describing a common event and involved objects with metaphor:** tail → thread; fly → sew;
If you take off my skin, I will not cry, but you will. What am I?	(A) grape (B) **onion (√)** (C) package (D) plant (E) **body (×)**	**Personalization.** Cutting onions → taking off my skin.
What is that which, though black itself, enlightens the world without burning?	(A) coal (B) hole (C) cd player (D) **sunlight (×)** (E) **ink (√)**	**Figure of speech** (ink → writing → knowledge → *light* of wisdom) **+ Counterfactual** (without burning)
I have hundreds of legs, but I can only lean. What am I?	(A) **chair (×)** (B) sock (C) pleopod (D) pants (E) **broom (√)**	**Counterfactual** (many legs but cannot stand) **+ Metaphor** (bristles)

Figure 5: Case studies of the error by UnifiedQA-3B model on the test set of RIDDLESENSE.

et al., 2018), pun generation (He et al., 2019; Luo et al., 2019), creative story generation, and humor detection (Weller and Seppi, 2019, 2020), sarcasm generation (Chakrabarty et al., 2020), etc.

Riddling, as a way to use creative descriptions to query a common concept, are relatively underexplored. Previous works (Tan et al., 2016; Gonçalo Oliveira and Rodrigues, 2018) focus on the generation of riddles in specific languages and usually rely on language-specific features (e.g., decomposing a Chinese character into multiple smaller pieces). There is few datasets or public resources for studying riddles as a reasoning task, to the best of our knowledge. The proposed RID-DLESENSE is among the very first works connecting commonsense reasoning and computational creative, and provides a large dataset to train and evaluate models for answering riddle questions.

6 Conclusion

We propose a novel commonsense reasoning challenge, RIDDLESENSE, which requires complex commonsense skills for reasoning about creative and counterfactual questions, coming with a large multiple-choice QA dataset. We systematically evaluate recent commonsense reasoning methods over the proposed RIDDLESENSE dataset, and find that the best model is still far behind human performance, suggesting that there is still much space for commonsense reasoning methods to improve. We hope RIDDLESENSE can serve as a benchmark dataset for future research targeting complex commonsense reasoning and computational creativity.

Ethical Considerations

Copyright of Riddles. The RiddleSense dataset is consistent with the terms of use of the fan websites and the intellectual property and privacy rights of the original sources. All of our riddles and answers are from fan websites that can be accessed freely. The website owners state that we may print and download material from the sites solely for *non-commercial use* provided that we agree not to change or delete any copyright or proprietary notices from the materials. Therefore, in addition to the dataset itself, we also provide the according copyright statements of every website and an URL link to the original page for each riddle. The dataset users must sign an informed consent form that they will only use our dataset for *research purposes* before they can access the both the riddles and our annotations.

References

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 LeBras, Jianfeng Gao, and Yejin Choi. 2020. 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.

Michael Boratko, Xiang Li, Tim O'Gorman, Rajarshi Das, Dan Le, and Andrew McCallum. 2020. ProtoQA: A question answering dataset for prototypical common-sense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1122–1136, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. 2020. R^3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7976–7986, Online. Association for Computational Linguistics.

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. Codah: An adversarially authored question-answer dataset for common sense. *ArXiv*, abs/1904.04365.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, A. Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. 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, Volume 1 (Long and Short Papers)*,
pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yanai Elazar, Abhijit Mahabal, Deepak Ramachandran, Tania Bedrax-Weiss, and Dan Roth. 2019. How large are lions? inducing distributions over quantitative attributes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3973–3983, Florence, Italy. Association for Computational Linguistics.

Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1295–1309, Online. Association for Computational Linguistics.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 607–613, Brussels, Belgium. Association for Computational Linguistics.

Hugo Gonçalo Oliveira and Ricardo Rodrigues. 2018. Exploring lexical-semantic knowledge in the generation of novel riddles in Portuguese. In *Proceedings of the 3rd Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2018)*, pages 17–25, Tilburg, the Netherlands. Association for Computational Linguistics.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30.

He He, Nanyun Peng, and Percy Liang. 2019. Pun generation with surprise. 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 1734–1744, Minneapolis, Minnesota. Association for Computational Linguistics.

Edward Hirsch. 2014. *A poet's glossary*. HMH.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. 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 2829–2839, Hong Kong, China. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020a. 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)*, pages 6862–6868, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Xiang Ren, and William Cohen. 2021. Differentiable open-ended commonsense reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4611–4625, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020b. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Fuli Luo, Shunyao Li, Pengcheng Yang, Lei Li, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. Pun-GAN: Generative adversarial network for pun generation. 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 3388–3393, Hong Kong, China. Association for Computational Linguistics.

Shangwen Lv, Daya Guo, Jingjing Xu, Duyu Tang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and Songlin Hu. 2020. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. 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 8449–8456. AAAI Press.

David D. McDonald and Federica Busa. 1994. On the creative use of language: The form of lexical resources. In *Proceedings of the Seventh International Workshop on Natural Language Generation*.

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. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

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.

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.

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.

Chuanqi Tan, Furu Wei, Li Dong, Weifeng Lv, and Ming Zhou. 2016. Solving and generating Chinese character riddles. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 846–855, Austin, Texas. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. 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 5998–6008.

Tony Veale. 2011. Creative language retrieval: A robust hybrid of information retrieval and linguistic creativity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 278–287, Portland, Oregon, USA. Association for Computational Linguistics.

Peifeng Wang, Nanyun Peng, Filip Ilievski, Pedro Szekely, and Xiang Ren. 2020. Connecting the dots: A knowledgeable path generator for commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4129–4140, Online. Association for Computational Linguistics.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh. 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 3621–3625, Hong Kong, China. Association for Computational Linguistics.

Orion Weller and Kevin Seppi. 2020. The rJokes dataset: a large scale humor collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6136–6141, Marseille, France. European Language Resources Association.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Frank F. Xu, Bill Yuchen Lin, and Kenny Zhu. 2018. Automatic extraction of commonsense LocatedNear knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 96–101, Melbourne, Australia. Association for Computational Linguistics.

Yichong Xu, Chenguang Zhu, Ruochen Xu, Yang Liu, Michael Zeng, and Xuedong Huang. 2020. Fusing context into knowledge graph for commonsense reasoning. *arXiv preprint arXiv:2012.04808*.

Jun Yan, Mrigank Raman, Tianyu Zhang, Ryan Rossi, Handong Zhao, Sungchul Kim, Nedim Lipka, and Xiang Ren. 2020. Learning contextualized knowledge structures for commonsense reasoning. *arXiv preprint arXiv:2010.12873*.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.